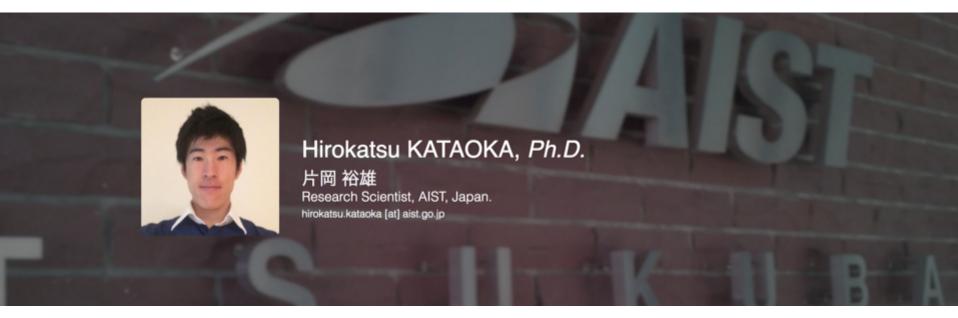
個人/グループサーベイ法

片岡 裕雄, Ph.D.

http://www.hirokatsukataoka.net/

片岡 裕雄(かたおかひろかつ)

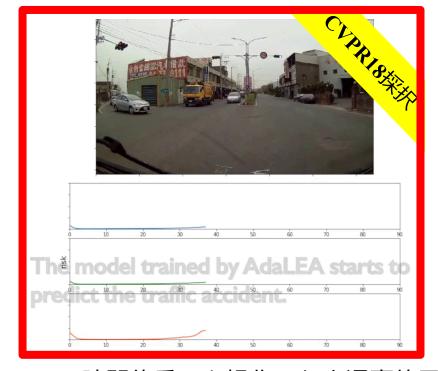
産業技術総合研究所 CV研究グループ/AL連携研究室/AIセンター 研究員 東京電機大学客員研究員



2014年慶應義塾大学大学院理工学研究科修了,博士(工学). 2013,2014年 ミュンヘン工科大学Visiting Scientist,2014年東京大学博士研究員,2015年産総研特別研究員.2016年4月より現職.画像認識,動画解析,人物行動解析に興味を持つ.cvpaper.challenge主宰.2011年ViEW小田原賞,2013年電気学会誌論文奨励賞,2014年藤原賞,2016年ECCV WS Brave New Idea Award.

mypage: http://www.hirokatsukataoka.net/

cvpaper.challenge: https://sites.google.com/site/cvpaperchallenge/





大規模交通二アミス動画収集

AdaLEA:時間的重みを操作した交通事故予測



文脈を弱教師とした異常検知



3次元畳み込みによる動画認識

片岡の主宰するcvpaper.challenge

論文読破・まとめ・発想・議論・実装・論文執筆・(社会実

装)に至るまで取り組むCVの今を映す挑戦

- 人員:産総研, 筑波大, 東京電機大, 慶應大による20名弱
- BraveNewなアイディアをトップ国際会議*に投稿

* Google Scholar Top-20にリストアップされている国際会議や論文誌

本取り組みの結果10本以上の論文 (含CVPRx2, ICRA, CVPRwx5, ECCVWx2, ICCVW) が採択 8件の招待講演,3件の国内外での受賞



HP, Twitter, SlideShareもご覧ください

HP: https://sites.google.com/site/cvpaperchallenge/

Twitter: @CVpaperChalleng SlideShare: @cvpaperchallenge

サーベイとは?

ひとことでいうと, その分野の動向を把握すること

- 現在どんな技術が流行っている?
- どういう歴史を辿ってきた?
- 自分のやっていることに最も近いものは?

II. RELATED WORK

A. Traffic data and approaches to its representation

Several practical databases for pedestrian detection, such as the French Institute for Research in Computer Science and Automation (INRIA) Dataset [3], Caltech [4], and the KITTI Vision Benchmark Suite for self-driving cars [2]) have been proposed in the past decade. The information contained in the KITTI database, which has been used to set meaningful vision problems for self-driving cars [2] as well as problems related to stereo vision, optical flow, visual odometry, semantic segmentation, two- and three-dimensional (2D/3D) object detection, and 2D/3D tracking, has proven especially useful.

In 2015, these problems were updated for stereo and optical flow [5]. Thanks to the development of sophisticated approaches, such as fully convolutional networks (FCN) [6] and region-based convolutional neural networks (R-CNN) [7], there has been improved performance of solving these problems using the KITTI benchmark dataset. In addition, a manner of geometry allows us to improve the rate of object detection [8] and optical flow [9] not only in stereo [10]. As for semantic segmentation, we can now obtain knowledge about dense connections and use this information with graphical models [11], [12].

Unfortunately, none of these datasets contain scenes of near-miss incidents in which pedestrians, cyclists, or other vehicles must be avoided. Thus, there is an urgent need for a collection of incident scenes that can be used to train self-driving cars on how to safely navigate such dangerous situations.

論文にもRelated workを書くこと多し

なぜ, サーベイをするのか?

トレンドの把握

- 知識がないと既存研究の劣化版を作りかねない
- トレンドを知らないと(天才でない限り)最先端の研究を生み出すことは 難しい

自身の研究の立ち位置を確認

- 何が違う?なぜやる?どこが良いのか?という哲学

究極的には次のトレンドを作るため(ここ重要)

- 分野の方向性を自ら定める
- より良く,正しい方向へ導く

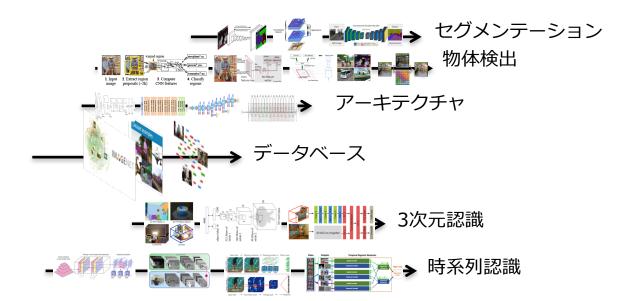
なぜ, サーベイするのか?

網羅的に調べることで分野の状況を精度よく映す

- サーベイの質や量が研究テーマ考案・選定に直結
- (グループ単位で) 1,000本/年は論文を読むポテンシャルがあると良い

技術の流れがわかる

- 点より線/面で技術を捉える(参考にするなら歴史を追う,他分野から学ぶ)
- 次に何をすれば良いかがなんとなくわかる



どれくらいサーベイしてるの?

個人/グループとしてサーベイを推進

- 「個人」で達成
 - 2015年度 615本, 2016年度 400⁺本
- 「グループ」で達成
 - CVPR 2015 完全読破(約10名)
 - 2016年, 1,000本読破達成(約20名)





意識的にサーベイして何が変わった?

2015年以前:個人プレー

- 論文調査:自分の狭い分野のみ

- 研究:従来法の単純な改善

組織的にサーベイしてから...

- 論文調査:網羅的に分野を把握

- 研究:物事の本質に迫るような問いを意識

- サーベイ/テーマ/実験に至るまで「質・効率を高める努力」を徹底
 - まだまだ徹底は不十分。。。

優れた問いを見つけよう!

新規に問題設定ができるようになった

- 思い付き(単純拡張)研究からの離脱
 - To invent, you need a good imagination and a pile of junk
 - 知識を詰め込むことで視野が繋がる感覚
- CVPR15完全読破から3年目, Top-tierの壁を突破
 - ICRA18採択! (Robotics Top-1)
 - CVPR18採択x2!! (CV&PR Top-1)

個人とグループの知識獲得がキモ

個人のサーベイ

グループのサーベイ

個人のサーベイ

グループのサーベイ

「分割と統合」の組み合わせが重要

個人サーベイで意識すること

速読と精読の組み合わせ

- 速読(50+本/月)
 - とにかく「広く浅く」
 - 研究テーマやアイディア考案の時に行うサーベイ法
- 精読(10-本/月)
 - 実装レベルで「狭く深く」
 - 具体的なテーマが決まっている際に

尺度を変えた読み方の統合で研究の効率化

論文を読もう!

論文は目的別に読むために体系化されている

- CVの論文を参考に解説します!

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

Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh National Institute of Advanced Industrial Science and Technology (AIST) Tsukuba, Ibaraki, Japan

The purpose of this study is to determine whether current video distancts how sufficient data for axising usery deep consolitational means interested (CNN), with quite-temporal drives disconsistent [10] heretich. Keently, the preformance interested and the second consolitation of the preformance interested and preformance interested and preformation are second hard only explored relatively shallow. But architectures or second new formation of the causine the architectures of union 200 (NN) from relatively shallow to sery deep ones on current who distances. Second on the result of those second new formation of the second control of the second new formation of the sec classions could be obtained: (1) ResNet-18 training r in significant overfitting for UCF-101, HMDB-51, and Acthingship has not for Electric. (a) The Electric almost has indigited and pair in milling of allow 20 CNN. (not almost has indigited and pair in milling of allow 20 CNN. (not been been been been almost a pair 13 Electrical states in the control of the control o witeNet but not for Kinetics. (ii) The Kinetics dataset has

(a) DCF-101 (selfc1). (b) HMDR-51 (selfc1).

are portided in this chance.

Activityble (1/2) provides samples from 200 human action classes with an overage of 137 unrimmed video per class and 141 activity issuances per video. Unlike the other dances, Activityble consists of unrimmed video, which include non-action frames. The nord video length is 3-90 hours, and the total number of action instances is 28,108. This chances is randomly upli nin them deflerent subsects training, validation, and returning. More specifically, 50% is used for validation, and 25% of the control of the contro

used for raining, 25% in used for validation, and 25% is, consistently as the contract of the

We began by training ResNet-18 on each dataset. Ac

we ogan by training research to an each trainer. Ac-cording to previous works [9, 16], 3D CNS trained on UCF-101, HMDB-51, and ActivityNet do not achieve high accuracy whereas ones trained on Kinetics work well. We tried to reproduce such results in this experiment. In this process, we used split 1 of UCF-101 and HMDB-51, and

4.1. Analyses of training on each dataset

ure 4: ResNet-18 training and validation looses. The validation looses on UCF-101, HMDB-51, and ActivityNet quickly or h values and were clearly higher than their corresponding training looses. The validation looses on Kinetics were slightly in corresponding training looses, significantly different turn those on the ether datasets.



creation of successful vision-based algorithms. In addition to such large-scale datasets, a large number of algorithms, such as residual learning [10], have been used to improve the control of the co

Figure 4 shows the training and validation losses esNet-18 on each dataset. As can be seen in the figure, didation losses on UCF-101, HMDB-51, and Activity

quickly converged to high values and were clearly higher

specially concepted to high values and were clearly higher cone that specifing promised with the translation of the other cone that specifing promised where the training on those three datasets. In addition to those looses, we confirmed pre-chip control to the control of the control of the control of the specific property of the control of the specific property of the control of the control of the control of the excursion in property of the control of the control of the excursion in property of the control of the control of the excursion in property of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the control of the excursion in the

Here, we will show ResNets accuracies changes based on model depths. Figure 2 shows the averaged accuracies over top-1 and top-5 ones. We can see that, essentially, as the depth increased, accuracies improved, and that the im-provements continued until reaching the depth of 152. We can also see that deeper ResNet-152 achieved significant im-

ement of accuracies compared with ResNet-18, which the previously examined architecture [9, 24]. In con-

icale of video datasets has begun to approach that of imag

For action recognition, CNNs with spatio-temporal threadmensional (20) convolutional bennel (5D CNNs) are recently more effective than CNNs with two-dimensional (20) kernels (21). From several years ago [14], 3D CNNS are explored to provide an effective tool for accurate action recognition. However, even the usage of well-orgalized models [23, 25] has failed to orecome the advantages of 2D-based CNNs that combine both stacked flow and RGB 2D-based CNNs that combine both stacked flow and RGB, magas [50]. The principal reasons for this faither has been the ranges [50]. The principal reasons for this faither has been the region of the principal reasons for the faither has been the first principal reasons and the state of the region of the reasons and the reasons and the reasons and the reasons and the reasons are the

However, can 3D CNNs retrace the successful history of 2D CNNs and ImageNet? More specifically, can the use of 3D CNNs trained on Kinetics produces significant progress in action recognition and other version used: The behavior in Figure 1.7 for Action with Pigure 1.7 for Action with Pigure 1.8 for the Figure 2.6 for Action with Pigure 2.6 for Action 1.7 for CNN should be a large-scale as interest particular to the Pigure 2.6 for Action 1.7 for Action 1.7

those datasets, we performed several experiments aimed at training and testing those architectures from scratch, as well

ResNet-200 started to overfit. Interestingly, the results are similar to those for 2D ResNets on ImageNet [11]. More specifically, the accuracies of both 2D and 3D ResNets im-proved as the depth increased utility caching the depth of 152, and then the accuracies did not increase when increas-ing the depth of 200. These results indicate that the Kinet-ies densed has sufficient data to train 3D CNNs in a measure

ble 2. Here, it can be seen that the accuracies of pe-activation ResNet-200 are slightly low when compared with the standard ResNet-200 though He et al. reported that the

Table 3 shows the results of the Kinetics test set used Table 3 shows the results of the Kinetics test set must to compare Backs(10), which acknowled the highest ac-curation, with the state-of-the-ort modulot. Here, it can be compared with CDD with both formatting [16], which is 10-layer network, as well as CDN-LSTM and two-steams CDN [16]. Then could also indicates the officiences of CDN [16]. Then could also indicates the officiences of DD trained on Kinetics from scrattle [1] were found to our perform Backs(2-CL) in the algorithm of the could be of perform Backs(2-CL) in the algorithm of the inter-pret may be compared to the could be of the countries of differences of the acrossic pass, "specifiedly, the state of the countries of the countries

ouncenoes or the network inputs. Specifically, the size of 13D is $3 \times 64 \times 224 \times 224$, whereas that of ResNeXt-101 is $3 \times 16 \times 112 \times 112$. Thus, 13D is 64 times larger than ResNeXt-101. To confirm the accuracies when using larger inputs, we also evaluated the ResNeXt-101 that used $3 \times 64 \times 112 \times 112$ inputs, which are the largest available sizes.

in our environment (NVIDIA TITAN X × 8). We can se



Figure 2: Averaged accuracies of 3D ResNets over top-1 and topon the Kinetics validation set. Accuracy levels improve as net depths increase. The improvements continued until reaching depth of 152. The accuracy of ResNet-200 is almost the sar that of ResNet-152. These results are similar to 2D ResNets on

as their fine-tuning. The results of those experiments (see Section 4 for details) show the Kinetics dataset can train 3D ResNet-152 from scratch to a level that is similar to the training accomplished by 2D ResNets on ImageNet, a es of future progress in action recognition and

training of very deep 3D CNNs from scratch for action recog-nition. Previous studies showed deeper 2D CNNs trained on ImageNet achieved better performance [10]. However, it is not trivial to show deeper 3D CNNs are better based on the previous studies because the data-scale of image datasets differs from that of video ones. The results of this study, which indicate deeper 3D CNNs are more effective, can be expected to facilitate further progress in computer vision for

2.1. Video Datasets

A couple of years after the abovementioned datasets wen introduced, larger video datasets were produced. These include ActivityNet [5], which contains 849 hours of video,

One of the popular approaches to CNN-based action necognition is the use of two-stream CNNs with 2D con-nectional kernels. In their study, Simonyan et al. proposed as method that uses Roll and stacked optical flow frames as appearance and motion information, respectively [20], and showed that combining the two-streams has the abtlty to improve action recognition accuracy. Since that tudy, numerous methods based on the two-stream CNNs study, uninerous methods based on the two-steam CNNs have been proposed to improve action recognition performance (6, 7, 8, 27, 28, 29). Utilité the showmentioned approaches, we focused on CNNs with 3D correductional learnets, which have recently began to comprehen 2D CNNs from an 2D CNNs from the performance video demonst. These 3D CNNs are intutively effective because such 3D convolution can be used to directly extract because such 3D convolution can be used to directly extract

The HMDB-51 [17] and UCF-101 [21] datasets are cur-rently the most successful in the field of action recogni-tion. These datasets gained significant popularity in the early years of the field, and are still used as popular benchmarks. However, a recent consensus has emerged that indicates tha they are simply not large enough for training deep CNNs

Method	Top-1	Top-5	Average
ResNet-18	54.2	78.1	66.1
ResNet-34	60.1	81.9	71.0
ResNet-50	61.3	83.1	72.2
ResNet-101	62.8	83.9	73.3
ResNet-152	63.0	84.4	73.7
ResNet-200	63.1	84.4	73.7
ResNet-200 (pre-act)	63.0	83.7	73.4
Wide ResNet-50	64.1	85.3	74.7
ResNeXt-101	65.1	85.7	75.4
DenseNet-121	59.7	81.9	70.8
DenseNet-201	61.3	83.3	72.3

0	
8	Table 3: Accuracies on the Kinetics test s
	accuracy over Top-1 and Top-5. Here, we n
	and Taxosteam ERD trained from scratch I

Method	Top-1	Top-5	Average
ResNeXt-101			74.5
ResNeXt-101 (64f)	-	-	78.4
CNN+LSTM [16]	57.0	79.0	68.0
Two-stream CNN [16]	61.0	81.3	71.2
C3D w/ BN [16]	56.1	79.5	67.8
RGB-I3D [3]	68.4	88.0	78.2
Two-stream [3D [3]	71.6	90.0	80.8

4.3. Analyses of fine-tuning Finally, in this section we confirm the performance of UCF-101 and HMIDB-51. The results of this experiment are important for determining whether the 3D CNNs are effective for other datasets. It should be noted that, in this experiment, fine-mining use only performed to train courts_3 and the fully councered layer became it achieved the best performance during the preliminary experiments.

Table 4 shows the accuracies of Kinetics pertained 3X.

Table 4 shows the accuracies of Kinetics pertained 3X.

Table 4 shows the accuracies of Kinetics pertained 3X.

Method	UCF-101	HMDB-5
ResNet-18 (scratch)	42.4	17.1
ResNet-18	84.4	56.4
ResNet-34	87.7	59.1
ResNet-50	89.3	61.0
ResNet-101	88.9	61.7
ResNet-152	89.6	62.4
ResNet-200	89.6	63.5
DenseNet-121	87.6	59.6
PacNaVe 101	60.7	618

rained ResNet-18 clearly outperformed one trained from cratch. This result indicate that pretraining on Kinetics is ffective on UCF-101 and HMDB-51. We can also see that

datasets. However, the OpensMet-121 results were hower than those of ResNet-0, thereby indicating this tay greater efficiency did not contribute on fine-tening of 3D this greater efficiency did not contribute on fine-tening of 3D this We shows the results of our comparison with state-of-the-art methods in Table 5. Hore, we can see that ResNet/ 101 achieved higher accuracies compared with COD [23], PDD [19], two-sneam CNS [28], and TDD [27]. Further-ison, we can also see that ResNet/Net [101] [27]. Further-tening, we can also see that ResNet/Net [101] [27]. Further-tening, we can also see that ResNet/Net [101] [27]. Further-tening than the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contributio

their Kitetics dataset were comparable with the results of 2D CNNs persisted on ImageNet: even though the results of 3D CNNs trained on the UCF101 and HMDB31 datasets were inferire to the 2D CNNs results in Bull another study. Carriera et al., proposed inception [22] based 3D CNNs, which they referred to a ED, and advised stance-of-the-art performance [2]. More recently, some works introduced Bersela extinctioners into 3D CNNs [9, 24], though they examined only relatively shallow ones.

3. Experimental configuration

duriests, a well as the UCF-101 and HMDIR-51 datasets. Other large datasets such as Spect-104 [15] and Yorlin-8-M [11] have been proposed. Although those years are supposed. Although those supposed in the s

In this study, in order to determine whether current video datasets have sufficient data for training of deep 3D CNNS, we conducted the three experiences described below us-ing UCF-101 [21], HMDB-51 [17], ActivityNet [5], and Kinetics [16]. We first examined the training of relatively studies 3D CNNS from scranch on early video danset. According to previous works [9, 16], 3D CNNs trained on UCF-101, HMDB-51, and ActivityNet do not achieve high CCT-101, HMMb-51, and Activitybed to not active high accuracy where ones transiend festiciates work well. We try to reproduce such results to accertain whether the datasets are sufficient data for deep 3D CNNS. Specifically, we used ResPert 18, which is the shallowest ResNet architec-ture, based on the assumption that if the ResNet 3E overfits when heig trained on a dataset, that diameter is too small to be used for training deep 3D CNNs from scranch. See Section 4.1 for details

we then conducted a separate experiment to observative whether the Kinetics dataset could train deeper 3D CNNs. A main point of this trial was to descriative how deeply the datasets could train 3D CNNs. Therefore, we trained 3D ResNets on Kinetics white varying the model depth from 18 to 200. If Kinetics can train vary deep CNNs, such as ResNet-152, which achieved the best performance in ResNets on ImageNet [10], we can be confident that they have sufficient data to train 3D CNNs. Therefore, the result of this experiment are expected to be very important for the future progress in action recognition and other video tasks. See Section 4.2 for details.

to action good performance oversoon small datasets, we ex-pect that the deep 3D ResNets pretrained on Kinetics would perform well on relatively small UCF-101 and HMDB-51. This experiment examines whether the transfer visual representations by deep 3D CNNs from one domain to another domain works effectively. See Section 4.3 for details.

Two-stream I3D [3] 3D 98.0 Two-stream CNN [20] 2D TDD [27] ST Multiplier Net [7] 2D

conv, 17, F

Figure 3: Block of each architecture. We represent cosy, x², F as the kernel size, and the number of feature maps of the co

Model Block	conv1	00	n/2_x	0	on/3_x		oes4_x	con	×5_x		
			F	N	F	N	F	N	F	N	
ResNet- (18, 34)	Basic		64	{2, 3}	128	{2,4}	256	(2, 6)	512	{2, 3}	
ResNet-(50, 101, 152, 200)	Bottleneck	₫ <u>-</u> ; , ,	64	3	128	{4, 4, 8, 24}	256	{6,23, 36,36}	512	3	pool.
Pre-act ResNet-200	Pre-act	×7×7,0 al stride 1 stride	64	3	128	24	256	36	512	3	global average pool, C-d fully-connected, softmax
WRN-50	Bottleneck		128	3	256	4	512	6	1024	3	100
ResNeXt-101	ResNeXt	spatial	128	3	256	24	512	36	1024	3	85
DenseNet- [121, 201]	DenseNet		64	{6, 6}	128	(12, 12)	256	{24, 48}	{512, 896}	{16, 32}	90

is one of the most successful architectures in image class

pievono studies that elemente only influent 2 in Rescue al-ritherature [9, 24], we examine not only desper architectures but also some extended versions of RenNer. Re particular, we explore the following architectures: ResNet (basis: and bottleneck blocks) [10], pre-activated ResNet [11], wide ResNet (WRN) [31], ResNext [30], and Demoshet [12]. The architectures are summarized in Figure 3 and Table 1. In the following peragraphs, we will briefly strondere each

A basic RosNets block consists of two completional la tional layer is followed by batch a nalization and a ReLU. A shortcut pass conn he block to the layer just before the last Rel.U in the block ResNet-18 and 34 adopt the basic blocks. We use identit

ResNot-18 and 34 adopt the basic blocks. We mee identity connections and zone pauling for the shortcost of the basic blocks (type A in [10]) to need increasing the sumber of basic blocks. The properties of these relatively-shallow reviews the countries of these cheen controlled the control of the properties of the controlled three controlled three particles of the controlled three particles of the school and 3 x x 3 x 3. The shortcut pass of this block is the same as that of the basic block. Robbects of this block is the same as that of the basic block. Robbects of this block is the same as the properties of the basic block in the same and the basic blocks. Robbects of this block is the same as the properties of the block of the basic blocks and the basic blocks. Robbects of the block is the same as the properties of the block in the same and the basic block is the same as the properties of the block in the same and the basic blocks are the block in the same and the same and the block in the same and th

a ReLU, whereas each batch norm activation ResNet is followed by the ReLU and a co

GPUs [31]. In this study, we evaluate the WRN-50 using a widening factor of two.

ResNeXt introduces cardinality, which is a different di-mension from deeper and wider. Unlike the original bot-tleneck block, the ResNeXt block introduces group convo-Cardinality refers to the number of middle conv ayer groups in the bottleneck block. In their study, Xie et

Dense/Net makes connections from early upgers to sure-papers by the use of a concentration that is different from the ResNets summation. This concatenation connects each layer densely in a Seed-forward fastion. Desse/Nets also adopt the per-activation used in pre-activation ResNets. In their study, Huang et al. showed that it achieves been accuracy and the property of the property of the property of the property of the many DesNets [12]. In this south was

3.3. Implementation Training. We use stochastic gradient descent with mome

num to train the networks and randomly spoarnes training samples from visions in training data in order to perform data augmentation. First, we select a temporal position in a video by uniform sampling in order to guestras attain-ing sample. A 16-frame clip is then guestrast attain-ing sample. A 16-frame clip is then guestrast around the selected emporal position. If the video is shorter than 16-frames, then we loop it as many times as necessary. Next, we randomly select a spatial position from the 4 corners or the center. In addition to the spatial position, we also select a spatial scate of the sample in order to perform multi-scale. cropping. The procedure used is the same as 1281. The channel. All generated samples retain the same class labels

m our thining, we use 'cross-energy sosses and unsa-propagate het gradhens. The training parameters include a weight decay of 0.001 and 0.0 for memorani. When crash 1, and other left by 10 for the validation is not sentence. When performing first ming, we start from a learning rate of 0.001, and seeing a weight decay of 1e-5. Recognition. We adopt the shifting window manner to gen-

inputs, but that the best performance could be achieved by combining RGB and optical flows. Meanwhile, Kay et al. showed that the results of 3D CNNs trained from scratch on

because such 3D convolution can be used to directly extract spatio-temporal features from raw videos. For example, 1st et al. proposed applying 3D convolution to extract spatio-temporal features from videos, while Tran et al. trained 3D

CNNs, which they referred to as C3D, using the Sports-

that using optical flows as inputs to 3D CNNs resulted in a higher level of performance than can be obtained from RGB

including 28,000 action instances. ActivityNet also provides some additional tacks, onto a merriment dessibilitation and detection, but the number of action instances is still on the other of tens of threasts. This year (2017), in an effort to ceate a successful pretrained model, Kay et al. ebased the Knetics datused [16]. The Knetics distanct includes more than 900,000 trimuned videos covering 400 categories. In order to determine whether it can train deeper 3D CNNs, we perfermed a number of experiments using these recent

e performed a number of experiments using these re-atasets, as well as the UCF-101 and HMDB-51 dataset

Method	UCF-101	HMDB-51
ResNet-18 (scratch)	42.4	17.1
ResNet-18	84.4	56.4
ResNet-34	87.7	59.1
ResNet-50	89.3	61.0
ResNet-101	88.9	61.7
ResNet-152	89.6	62.4
ResNet-200	89.6	63.5
DenseNet-121	87.6	59.6
ResNeXt-101	50.7	63.8

5. Conclusion

In the sindy, we assumed the architectures of various CNNs with spatio-empreed 3D consolitated laterate on CNNs with spatio-empreed 3D consolitated laterates on CNNs with spatio-empreed 3D consolitates confirming for engineerings, the filtering conclusions could be refused on the spatial of spatial concerning for the CNNs of the CNNs of

geNet experienced significant progress in various tasks such as object detection, semantic segmentation, and image cap-tioning. It is felt that, similar to these, 3D CNNs and Kinet-ics have the potential to contribute to significant progress in as have the potential to contribute to significant progress in fields related to various video tasks such as action detection video summarization, and optical flow estimation. In our future work, we will investigate transfer learning not only for action recognition but also for other such tasks.

- Reterences

 11] S. Abrol-Delga, N. Kethart, J. Lee, P. Nators, G. Toderici,
 B. Virradurgia, and S. Viginaruminthan, YouThe-SM: A large-scale video coloridation benefition. Act To approach (1998) and A. Tolerman. Quo vadio, action recognition? A new model and the Kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni-Cal Conference on Computer Vision and Pattern Recogni-Cal Conference on Computer Vision and Pattern Recogni-Cal Conference on Computer Vision and Pattern Recogni-

- mon's how model and the Scottics affects and Experimental

 1. Deep, W. Deep, S. Scotte, L. J. B., K. L. and I. Feifer,
 Insignifies, A. Lago Scient Estimation lamp Database. In

 John S. L. Barris, S. Scotte, L. J. B., K. L. and I. Feifer,
 Insignifies, A. Lago Scient Estimation lamp Database. In

 John S. Congolius C. Scott, C. Scott, C. Scott, C. Scott,
 J. B. G. Datas C. Sch Bellense, Virge Feneral and J. C.

 1. B. G. Datas C. Sch Bellense, Virge Feneral and J. C.

 1. B. G. Datas C. Sch Bellense, Virge Feneral and J. C.

 Johnson Secophism (Exp. Prince-deep of the MEEI ConJourney on Company Training and Prince-Scotting of MEEI ConJourney on Company Training and Prince
 Journey on Co

- Erneties Recognition, 2017, 2, 3, 4, 6
 [10] K. He, X. Zhang, S. Ren, and J. Sun. Deep evidual learning for trapp recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 710–778, 2016, 1, 2, 3, 4, 5
 [11] K. He, X. Zhang, S. Ren, and J. Sun. Mentity mappings in deep residual networks. In Proceedings of the Environment
- Josepheroce on Long-2016. 2, 4, 5, 7 G. Huang, Z. Liu, L. van der Masten, and K. Q. Weinberger Connected convolutional networks. In Proceedings Densely contented convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Flattom Recognition (CVPR), pages 4700-4716, 2017, 2, 4, 5, 7
 [3] S. Jolfe and C. Sergody. Both commitmation: Accelerating deep network training by reducing internal convariate shiftly. In Proceedings of the International Conference on Many of the Membershaled Conference on Many of the Membershaled Conference on Many of the International Con

dataset of 101 human action closes from videos in the wild. CRCV-WR-10-10, 2012. 1, 2, 3, 5 22. C. Steeple, W. Lu, Y. Jiu, P. Sermanet, S. Reed, D. Angueles, D. Erfins, V. Valanskie, and A. Rabierski. Goily dever-sited control time. In Proceedings of the IEEE Conference on Computer Vision and Fattern Recognition (CVPR), pages 30 – 40, 2015. The Conference on Computer Vision and Fattern Recognition (CVPR), pages 30 – 30, 2015. The Conference on Computer Vision and Pattern Recognition of CVPR, pages 30 – 30, 2015. The Conference on Computer Vision and Pattern Recognition of CVPR, pages 30 – 30, 2015. The Conference on Computer Vision and Computer Vision of CVPR, pages 30 – 30, 2015. The Conference on Computer Vision of CVPR, pages 30, 2015. The Conference on Computer Vision of CVPR, pages 30, 2015. The Conference on CVPR pages 30, 2015. The Conference on CVPR pages 30, 2015. The CVP

[25] D. Limit, Li. Seesaway and Control States with 3D consolutional na-learing spatiateraporal features with 3D consolutional na-works. In Proceedings of the International Conference on Camputer Vision (ICCV), pages 4480–4497, 2015. 2, 8
[24] D. Tran, J. Ray, Z. Shen, S. Chang, and M. Polart. Convent architecture search for spatiotemporal feature learning. arXiv: 2017.01.2017, 2017. 2, 1, 4

79, 2013. 6
127] L. Wang, Y. Qiao, and X. Tang. Action recognition with

気をつけていること

森を見る,木を見る

- 森:ざっと全体を通して見る
 - 速読ならこれで十分
 - 背景, イントロ, 図表, 結果, 結論を中心に
- 木:細かいところまで目を通す (但し, 目的を見失わない)
 - どんな情報が欲しいかを明確にして読む
 - 実装したい?輪講資料を作りたい?

アブストラクトを見る

– 良いアブストラクトだとなんとなく全体がわかる

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

Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh National Institute of Advanced Industrial Science and Technology (AIST) Tsukuba, Ibaraki, Japan

{kensho.hara, hirokatsu.kataoka, yu.satou}@aist.go.jp

Abstract

The purpose of this study is to determine whether current video datasets have sufficient data for training very deep convolutional neural networks (CNNs) with spatio-temporal three-dimensional (3D) kernels. Recently, the performance levels of 3D CNNs in the field of action recognition have improved significantly. However to date conventional research has only explored relatively shallow 3D architectures. We examine the architectures of various 3D CNNs from relatively shallow to very deep ones on current video datasets. Based on the results of those experiments, the following conclusions could be obtained: (i) ResNet-18 training resulted in significant overfitting for UCF-101, HMDB-51, and ActivityNet but not for Kinetics. (ii) The Kinetics dataset has sufficient data for training of deep 3D CNNs, and enables training of up to 152 ResNets layers, interestingly similar to 2D ResNets on ImageNet. ResNeXt-101 achieved 78.4% average accuracy on the Kinetics test set. (iii) Kinetics pre trained simple 3D architectures outperforms complex 2D ar chitectures, and the pretrained ResNeXt-101 achieved 94.5% and 70.2% on UCF-101 and HMDB-51, respectively. The use of 2D CNNs trained on ImageNet has produced significant progress in various tasks in image. We believe that using deep 3D CNNs together with Kinetics will retrace the successful history of 2D CNNs and ImageNet, and stimulate advances in computer vision for videos. The codes and pretrained models used in this study are publicly available.

1. Introduction

The use of large-scale datasets is extremely important when using deep convolutional neural networks (CNNs), which have massive parameter numbers, and the use of CNNs in the field of computer vision has expanded significantly in recent years. ImageNet [4], which includes more

https://github.com/kenshohara/3D-ResNets-PyTorch

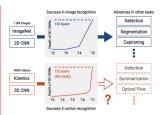


Figure 1: Recent advances in computer vision for images (top) and videos (bottom). The use of very deep 2D CNNs trained on ImageNet generates outstanding progress in image recognition as well as in various other tasks. Can the use of 3D CNNs trained on Kinetics generates similar progress in computer vision for videos?

than a million images, has contributed substantially to the creation of successful vision-based algorithms. In addition to such large-scale datasets, a large number of algorithms, such as residual learning [10], have been used to improve image classification performance by adding increased depth to CNNs, and the use of very deep CNNs trained on ImageNet have facilitated the acquisition of generic feature representation. Using such feature representation, in turn, has significantly improved the performance of several other tasks including object detection, semantic segmentation, and image captioning (see top row in Figure 1).

To date, the video datasets available for action recognition have been relatively small when compared with image recognition datasets. Representative video datasets, such as UCF-101 [21] and HMDB-51 [17], can be used to provide realistic videos with sizes around 10 K, but even though they are still used as standard benchmarks, such datasets are obviously too small to be used for optimizing CNN representations from scratch. In the last couple of years, ActivityNet [5], which including 28,000 action instances. ActivityNet also provides some additional tacks, onto a merriment dessibilitation and detection, but the number of action instances is still on the other of tens of threasts. This year (2017), in an effort to ceate a successful pretrained model, Kay et al. ebased the Knetics datused [16]. The Knetics distanct includes more than 900,000 trimuned videos covering 400 categories. In order to determine whether it can train deeper 3D CNNs, we perfermed a number of experiments using these recent e performed a number of experiments using mese re-stasets, as well as the UCF-101 and HMDR-51 dataset

duriests, a well as the UCF-101 and HMDIR-51 datasets. Other large datasets such as Spect-104 [15] and Yorlin-8-M [11] have been proposed. Although those years are supposed. Although those supposed in the s

2.2. Action Recognition Approaches

One of the popular approaches to CNN-based action necognition is the use of two-stream CNNs with 2D con-nectional kernels. In their study, Simonyan et al. proposed as method that uses Roll and stacked optical flow frames as appearance and motion information, respectively [20], and showed that combining the two-streams has the abtl-

sally, misroes methods based on the two-serim CNNs have been proposed a programme (Ex. E. 27, 27, 28, 28).

Consider the proposed the recognition performance (Ex. E. 27, 28, 28, 28).

CONS with 3D connectional larends, which have neverthybegue to comprehensing DCNNs through time of Engress-collection and the contract of the contract

ough for training deep CNNs

Its are similar to 2D ResNets or

sults of those experiments (so the Kinetics dataset can train th to a level that is similar to

w 2D ResNets on ImageNet, a

n those results, we will discu

ress in action recognition an

from scratch for action recog od deeper 2D CNNs trained

	Top-1	Top-5	Average
	54.2	78.1	66.1
	60.1	81.9	71.0
	61.3	83.1	72.2
	62.8	83.9	73.3
	63.0	84.4	73.7
	63.1	84.4	73.7
30	63.0	83.7	73.4
	64.1	85.3	74.7
	65.1	85.7	75.4
	59.7	81.9	70.8
	61.3	83.3	72.3

Kinetics test set. Assunge is average p-5. Here, we refer the results of RGI d from scratch [3] for fair comparison					
	raten (Ton-5	Averan		
	op-1	rop-3	Average		
	-	-	74.5		
	-	-	78.4		
	57.0	79.0	68.0		
	61.0	81.3	71.2		
	56.1	79.5	67.8		
	68.4	88.0	78.2		
	71.6	90.0	80.8		

ty to improve action recognition accuracy. Since that tudy, numerous methods based on the two-stream CNNs

de facio studind for 3D CNNs. They also experimentally doubt along 34 x 32 correctionnal transit achieved the best performance level. In another study, Varuel et al. showed that expanding the temporal length of spits for COI improves exceptation performance [25]. Those authors also found that using spitted flows as a pagin so 10 CNNs resulted in a higher level of performance han can be obtained from ROII constitution [36]. The second performance has can be contained from ROII constituting ROII and optical flows. Meansthin Koy et al. showed that the results of 3D CNNs trainfed from scratch on

Table 4: Top-1 accuracies on UCF-101 and HMDB-51. All accuracies are averaged over three solits.

Method	UCF-101	HMDB-5
ResNet-18 (scratch)	42.4	17.1
ResNet-18	84.4	56.4
ResNet-34	87.7	59.1
ResNet-50	89.3	61.0
ResNet-101	88.9	61.7
ResNet-152	89.6	62.4
ResNet-200	89.6	63.5
DenseNet-121	87.6	59.6
ResNeXt-101	50.7	63.8

trained ResNet-18 clearly outperformed one trained from scratch. This result indicate that pretraining on Kinetics is effective on UCF-101 and HMDB-51. We can also see that effective on UCF-101 and HMDB-51. We can also see that the accuracies bestudely improved as the depth increased, similar to the results obtained on Kinetics. However, un-like the results on Kinetics, Resbelt-20 also improved the accuracies in HMDB-51. Because, as described above, the fine-tuning in this experiment was only performed to train corn²5, and the fully connected layer, the numbers of trained parameters were the same from Resbelt-50 in Resbelt-200. Therefore, the pretrained early layers, which may be a superior of the results of the results of the results of the major of the results of the results of the results of the results of the major of the results of the results of the results of the results of the major of the results of the results of the results of the results of the major of the results of the results of the results of the results of the major of the results of the results of the results of the results of the major of the results of the result

school od. In Selection, for permanel couly inject, visited manages. These revolutions that for former representations of Roches 200 would be entailed for HIGHES 51 cent indicates the result of the selection of Roches 200 would be entailed for HIGHES 51 cent indicates the least of the result of We shows the results of our comparison with state-of-the-art methods in Table 5. Here, we can see that ResNeXt-1101 (1997) and the shewed higher accuracies compared with C3D [23], P3D [19], two-stream CNN [20], and ITDD [27]. Further-more, we can also see that ResNeXt-101 (64f), which utilize

2D CNNs pretrained on ImageNet, even though the results of 3D CNNs trained on the UCF101 and HMDB51 datasets

3. Experimental configuration

In this study, in order to determine whether current video datasets have sufficient data for training of deep 3D CNNS, we conducted the three experiences described below us-ing UCF-101 [21], HMDB-51 [17], ActivityNet [5], and Kinetics [16]. We first examined the training of relatively studies 3D CNNS from scranch on early video danset. According to previous works [9, 16], 3D CNNs trained on UCF-101, HMDB-51, and ActivityNet do not achieve high CCF-101, HSMD8-51, and ActivityNet do not dathere light accuracy whereones transien of kittens wat with the con-curacy whereones transien of kittens wat with the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the contraction of the contraction of the con-traction of the contraction of the Section 4.1 for details.

We then conducted a separate experiment to determine whether the Kinetics dataset could wain deeper 3D CNNs. A main point of this will us to determine how deeply the datasets could train 3D CNNs. Therefore, we trained 3D Res/Nest on Kinetics while varying the model depth from 18 to 20.0 If Kinetics can train were deep CNNs, such as Res/Nest-152, which achieved the best performance in as ResNet-152, which achieved the best perfortance in ResNets on ImageNet [10], we can be confident that they have sufficient data to train 3D CNNs. Therefore, the results of this experiment are expected to be very important for the future progress in action recognition and other video tasks. See Section 4.2 for details.

In the final experiment, we examined the fine-tuning or kinetics pretrained 3D CNNs on UCF-101 and HMDB-51 since pretraining on large-scale discisses is an effective way to achieve good performance levels on small disasses, we exto achieve good performance levels on small datasets, we ex-pect that the deep 3D ResNets pertained on Kinetics would perform well on relatively small UCF-101 and HMDB-51. This experiment examines whether the transfer visual rep-resentations by deep 3D CNNs from one domain to another domain works effectively. See Section 4.3 for details.

with 3D conformations used in this study, moreover, minests one of the most successful architectures in image classification, provides shortcut connections that allow a signa to bypass one layer and move to the next layer in the se

- Ernstein Recipation, 2017, 2, 3, 4, 6
 [10] K. He, X. Zhang, S. Rei, and J. Sin. Deep residual learning
 for image recognition. In Proceedings of the IEEE Conference
 on Company Video and Patrice Recognition (CVPR), pages,
 770–778, 2016, 1, 2, 3, 4, 5
 [11] K. He, X. Zhang, S. Best, and J. Sen. Meetily mappings
 in deep residual authority. In Proceedings of the European
 (in Computer Vision (ECCV), pages 609–645,
 2016, 2, 4, 5
 (Computer Vision (ECCV), pages 609–645,
 2016, 2, 4, 5
- Densyl contexted convolutional networks. In Proceedings of the IEEE Conference or Computer Vision and Fattern Recognition (CVPR), pages 4700-4108, 2017, 2, 4, 5, 7

 [13] S. Heft and C. Stragell, Baths of committation: Accelerating deep network training by reducing internal convariate shifts. In Proceedings of the International Conference on Many of the Internation

A basic RosNets block consists of two completional la A fine face of the color to the color of the tional layer is followed by batch

tional lawer is followed by batch normalization and a ReLU, whereas each batch norm activation ResNet is followed by the ReLU and a co

evaluated.

The WRN architecture is the same as the ResNet (bost-tieneck), but there are differences in the number of feature maps for each corroditional layer. WRN increases the number of Sature maps rather than the number of Juyers. Such wide architectures are efficient in parallel computing using GPUs [31]. In this souty, we evaluate the WRN-50 using a widering factor of two.

widening factor of two.

ResNeX1 introduces cardinality, which is a different dimersion from deeper and wider. Utilike the original bottleneck block, the ResNeXt block introduces group convolutions, which divide the feature maps into small groups. Cardinality refers to the number of middle convolutional layer groups in the bottleneck block. In their study, Xie et

3.3. Implementation

acilitates optimization in the training and reduces over-liting [11]. In this study, pre-activation ResNet-200 was

layer groups in the bottlework book. In their study, Xia or and, a showed that increasing the centralisting of Desentherizers is more effective than using water or deeper one [10]. In the study, we exhaust Res-Nock.101 using the centralisting that study, we exhaust Res-Nock.101 using the centralisting properties of the contraction of

Training. We use stochastic gradient descent with morn

um to trait the networks and randowly generate training samples from videos in training data in order to perform data augmentation. First, we select a temporal position in a video by uniform amplies in order to agreeate a train-saction of the contract of the contract of the contract selected temporal position. If the video is otherer than 16 of firms, show we hope its many times as accourage, beautiful entraction of the contract of the contract of the contract of the corner. In addition to the quality position from the 4 corners or a uputal scale of the sample in order to perform multi-scale copping. The precoducts used is the same at 13th. The und the single-would and engin are see some as the Snort-side length of the frame, and scale O.5 means that the sample is half the size of the sites riske length. The sample aspect ratio is I and the sample is segotis-temporally corpoped at the positions, scale, and aspect ratio. We spitially sestie the sample at 112×112 pitests. The size of each sample is 3 channels × 16 frames × 112 pitests, and each sample is horizontally flipped with 50% peochability. We also perform mean subteaction, which means that we subtract the nean values of ActivityNet from the sample for each colo channel. All generated samples retain the same class labels

In our training, we use cross-entropy losses and back m out thating, we use 'Cross-eatings' assess data teast-propagate their gradients. The training parameters include a weight decay of 0.001 and 0.0 for mentionism. When cratted 1, and observe which the contraction of the cratted 1, and observe which the contractions to the cratter of 0.001, and seep as weight decay of lee 5. Recognition. We adopt the bilding window manner to gra-responding to the contraction of the second of the contraction of t

Recognition. We adopt the sliding window anamer to gest-erate input (lap. (i.e., each whole is 1981 into non-overlapped 16-frame clips), and recognize actions in videos using the tazined networks. Each (lip is spatially) cropped around a center position at scale. I. We then input each clip into the networks and estimate the clip (to see, soccess, which are aver-aged over all the clips of the video. The class that has the maximum score indicates the recognized class label.)

DenseNet-[121, 201] DenseNet

- References

 11. S. Ano B-Halja, N. Kothart, J. Leo, P. Natser, G. Tokares,
 B. Vinnkarajas, and S. Vijaparamirahas. Teel flow-88-fs. Collisions,
 B. Serial States, and S. Vijaparamirahas. Teel flow-88-fs. And Teeprofest.
 Berges-sche video calculation between durft proposed,
 and the proposed control of the proposed collisions, and the proposed collisions and the proposed collisions.

 2011 S. Corter and A. Ziscortana, Cup valid, and the collisions and the proposed collisions. The proposed collisions are also proposed collisions and the proposed collisions. The proposed collisions are also proposed collisions are also proposed collisions. The proposed collisions are also proposed collisions are also proposed collisions. The proposed collisions are also proposed collisions are also proposed collisions. The proposed collisions are also proposed collisions are also proposed collisions. The proposed collisions are also proposed collisions are also proposed collisions. The proposed collisions are also proposed collisions are also proposed collisions. The proposed collisions are also proposed collisions are also proposed c
 - - - dataset of 101 human action classes, from videos in the wild. CRCV-TR, 12-01, 2012, 12, 23, 35

 C. Szegedy, W. Liu, Y. Jiu, P. Sermanet, S. Reed, D. Anguelov, D. Erfan, V. Vanhoueke, and A. Rabbosskin. Going deeper with currentitations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9, 2015, 3
 - [25] D. Limit, Li. Seesaway and Control States with 3D consolutional na-learing spatiateraporal features with 3D consolutional na-works. In Proceedings of the International Conference on Camputer Vision (ICCV), pages 4480–4497, 2015. 2, 8
 [24] D. Tran, J. Ray, Z. Shen, S. Chang, and M. Polart. Convent architecture search for spatiotemporal feature learning. arXiv: 2017.01.2017, 2017. 2, 1, 4

d action classes and instances.

We resized the videos to heights of 240 pixels without bancing their associated and then stored them.

Since the above mentioned experiment

4. Results and discussion

4.1. Analyses of training on each dataset We begin by training RosNot-18 on each dataset. Ac-cording to previous works 19, 163, 3D CNNs trained on UCF-101, HMD8-51, and ActivityNet do not achieve high accuracy whereas ones trained on Kinetics work well. We inted to reproduce such results in this experiment. In this process, we used split 1 of UCF-101 and HMD8-51, and

ment of accuracies compared with ResNet-18, which he previously examined architecture [9, 24]. In con-

in our environment (NVIDIA TITAN X × 8). We can se

perform ResNeXI-101 even though ResNeXI-101 is a deeper architecture than I3D. One of the reasons for this is the size differences of the ensewed rapus. Specifically, the size of 13D is $3 \times 64 \times 234 \times 224$, whereas that of ResNeXI-101 is $3 \times 16 \times 112 \times 112$. Thus, I3D is 64 times larger than ResNeXI-101. To confirm the accuracies when using larger inputs, we also evaluated the ResNeXI-101 that used 3.648 112 k12 Hzgus, which are the larger attaillab sixes $3.648 \times 112 \times 112$ Hzgus, which are the larger attaillab sixes and

UCF-101 and HMDB-51. The results of this experiment are important for determining whether the 3D CNNs are effective for other datasets. It should be noted that, in this ex-periment, fine-ratining wave only performed to train courts_3 and the fully connected layer because it achieved the best performance during the preliminary experiments. Table 4 shows the accuracies of Kinetics pretrained 33. CNNs, as well as Reselvel-18 trained from scratch, in UCF-

Finally, in this section we confirm the perform

Method	Dim	UCF-101
ResNeXt-101	3D	90.7
ResNeXt-101 (64f)	3D	94.5
C3D [23]	3D	82.3
P3D [19]	3D	88.6
Two-stream I3D [3]	3D	98,0
Two-stream CNN [20]	2D	88.0
TDD [27]	2D	90.3

that two-stream I3D [3], which utilizes simple two-stream

5. Conclusion

In the sindy, we assumed the architectures of various CNNs with spatio-empreed 3D consolitated laterate on CNNs with spatio-empreed 3D consolitated laterates on CNNs with spatio-empreed 3D consolitates confirming for engineerings, the filtering conclusions could be refused on the spatial of spatial concerning for the CNNs of the CNNs of

geNet experienced significant progress in various tasks such as object detection, semantic segmentation, and image cap-tioning. It is felt that, similar to these, 3D CNNs and Kinetfields related to various video tasks such as action detection video summarization, and optical flow estimation. In our future work, we will investigate transfer learning not only

04 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1

Figure 3: Block of each architecture. We represent coss, x³, F as the kernel size, and the number of feature maps of the co

 $A.3 \times 3 \times 3$ max-pooting layer (stride 2) is also located before conv2.x of all networks for down-sampling. In addition, conv1 spatially down-samples inputs with a spatial stride of two. C of the fully-connected layer is the number of classes.

| 10.8-x1 | 10.8-x1 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.

64 [6,6] 128 [12,12] 256 [24,48]

- (d) Deep W. Deep R. Sechee L. J. J. K. J. Land I. Tes Fre-land Deep R. Sechee L. J. J. K. J. Land I. Tes Fre-langsber A. Lang-Scale Bienerskies I lang Brishness. In Proceedings of the 2022 Confession on Confession May 16, 15 B. G. Dates C. Dea Biellows, View Beestra and J. C. J. C. Beltas C. Dea Biellows, View Beestra and J. C. J. Cheles, Assirphys. A lapsocal visit boundards for tea-turing the Confession of the Confession of the Confession of Confession and Confession of Computer View and Pattern Resignation (CVPR), pages 403–407, 2013. 1, 2, 3, 5, 6
 [6] C. Esternetzeter, A. Drez, and R. Willes, Spatian quality and Confession of Confession of Confession of Confession (CVPR), pages 403–407, 2013. 1, 2, 3, 5, 6
- umpter networks for vadeo action recognition. In Proceed-ge of the IEEE Conference on Computer Vision and Fastern recognition (CVPR), 2017. 5,8 Feithrenhofer, A. Pinz, and A. Zisserman. Convolutional re-stream activers: fusion for video action recognition. In

- Julie 2, 4, 5, 7

 3. Huang, Z. Liu, L. van der Masten, and K. Q. Weinberger

 and an annual convolutional networks. In Proceeding:

previous studies that examined only limited 3D Resched architectures [9, 24], we estamline not only deeper architectures but also some extended versions of RenSe. Red limited the studies of the studies of RenSe. Red limited the studies of RenSe. Red limited the studies and bottleneck blocks) [10], pre-activation RenSe [11], wide RegSel (WRN) [31], RenSeNEX [10], and DemoNet [11]. The architectures are summarized in Figure 3 and Table 1. In the following practipation we will briefly studied ceach in the following congruptions we will briefly studied ceach the following the studies of t

{16, 32}

[17] H. Kuchne, H. Ruang, E. Garrote, T. Poggio, and T. Serre

- 79, 2013. 6 [27] L. Wang, Y. Qiao, and X. Tang. Action recognition will

イントロ, 概念図を見る

- どんな問題設定?その研究の新規性は?

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

Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh National Institute of Advanced Industrial Science and Technology (AIST) Tsukuba, Ibaraki, Japan

{kensho.hara, hirokatsu.kataoka, yu.satou}@aist.go.jp

Abstract

The purpose of this study is to determine whether current video datasets have sufficient data for training very deep convolutional neural networks (CNNs) with spatio-temporal three-dimensional (3D) kernels. Recently, the performance levels of 3D CNNs in the field of action recognition have improved significantly. However to date conventional research has only explored relatively shallow 3D architectures. We examine the architectures of various 3D CNNs from relatively shallow to very deep ones on current video datasets. Based on the results of those experiments, the following conclusions could be obtained: (i) ResNet-18 training resulted in significant overfitting for UCF-101, HMDB-51, and ActivityNet but not for Kinetics. (ii) The Kinetics dataset has sufficient data for training of deep 3D CNNs, and enables training of up to 152 ResNets layers, interestingly similar to 2D ResNets on ImageNet. ResNeXt-101 achieved 78.4% average accuracy on the Kinetics test set. (iii) Kinetics pretrained simple 3D architectures outperforms complex 2D ar chitectures, and the pretrained ResNeXt-101 achieved 94.5% and 70.2% on UCF-101 and HMDB-51, respectively. The use of 2D CNNs trained on ImageNet has produced significant progress in various tasks in image. We believe that using deep 3D CNNs together with Kinetics will retrace the successful history of 2D CNNs and ImageNet, and stimulate advances in computer vision for videos. The codes and pretrained models used in this study are publicly available!

1. Introduction

The use of large-scale datasets is extremely important when using deep convolutional neural networks (CNNs). which have massive parameter numbers, and the use of CNNs in the field of computer vision has expanded significantly in recent years. ImageNet [4], which includes more

https://github.com/kenshohara/3D-ResNets-PvTorch

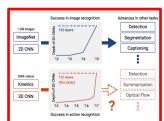


Figure 1: Recent advances in computer vision for images (top) and videos (bottom). The use of very deep 2D CNNs trained on ImageNet generates outstanding progress in image recognition as well as in various other tasks. Can the use of 3D CNNs trained on Kinetics generates similar progress in computer vision for videos?

than a million images, has contributed substantially to the creation of successful vision-based algorithms. In addition to such large-scale datasets, a large number of algorithms. such as residual learning [10], have been used to improve image classification performance by adding increased depth to CNNs, and the use of very deep CNNs trained on ImageNet have facilitated the acquisition of generic feature representation. Using such feature representation, in turn has significantly improved the performance of several other tasks including object detection, semantic segmentation, and image captioning (see top row in Figure 1).

To date, the video datasets available for action recognition have been relatively small when compared with image recognition datasets. Representative video datasets, such as UCF-101 [21] and HMDB-51 [17], can be used to provide realistic videos with sizes around 10 K, but even though they are still used as standard benchmarks, such datasets are obviously too small to be used for optimizing CNN representations from scratch. In the last couple of years, ActivityNet [5], which including 25.000 action instances. ActivityNet also provides some additional tests, such as untrimmed classification and detection, but the number of action instances is still on the order of these of thousands. This year (2017), in an effort in cotata a successful pretrained model, Kay et al. obtased the Kenties datused [16]. The Kinneiss distent includes more than 900,000 trummed videos covering 400 categories. In order to determine whether it can turn the obeyer 30 CNNs, we performed a number of CPP of the order of the obeyer and the obeyer 30 CNNs, we performed a number of CPP of the order of the obeyer 30 CNNs.

duriests, a well as the UCF-101 and HMDIR-51 datasets. Other large datasets such as Spect-104 [15] and Yorlin-8-M [11] have been proposed. Although those years are supposed. Although those supposed in the s

2.2. Action Recognition Approaches

One of the popular approaches to CNN-based action necognition is the use of two-stream CNNs with 2D con-nectional kernels. In their study, Simonyan et al. proposed as method that uses Roll and stacked optical flow frames as appearance and motion information, respectively [20], and showed that combining the two-streams has the abtlty to improve action recognition accuracy. Since that tudy, numerous methods based on the two-stream CNNs

sally, misroes methods based on the two-serim CNNs have been proposed a programme (Ex. E. 27, 27, 28, 28).

Consider the proposed the recognition performance (Ex. E. 27, 28, 28, 28).

CONS with 3D connectional larends, which have neverthybegue to comprehensing DCNNs through time of Engress-collection and the contract of the contract de facio studind for 3D CNNs. They also experimentally doubt along 34 x 32 correctionnal transit achieved the best performance level. In another study, Varuel et al. showed that expanding the temporal length of spits for COI improves exceptation performance [25]. Those authors also found that using spitted flows as a pagin so 10 CNNs resulted in a higher level of performance han can be obtained from ROII constitution [36]. The second performance has can be contained from ROII constituting ROII and optical flows. Meansthin Koy et al. showed that the results of 3D CNNs trainfed from scratch on

Method

ResNet-18 (scrans

ough for training deep CNN

ts are similar to 2D ResNets or

sults of those experiments (so the Kinetics dataset can train th to a level that is similar to

y 2D ResNets on ImageNet, a

those results, we will discus

ress in action recognition an

from scratch for action recog od deeper 2D CNNs trainer

	Top-1	Top-5	Average
	54.2	78.1	66.1
	60.1	81.9	71.0
	61.3	83.1	72.2
	62.8	83.9	73.3
	63.0	84.4	73.7
	63.1	84.4	73.7
30	63.0	83.7	73.4
	64.1	85.3	74.7
	65.1	85.7	75.4
	59.7	81.9	70.8
	61.3	83.3	72.3

Top-1	Top-5	Average
-	-	74.5
-	-	78.4
57.0	79.0	68.0
61.0	81.3	71.2
56.1	79.5	67.8
68.4	88.0	78.2
71.6	90.0	80.8

Table 4: Top-1 accuracies on UCF-101 and HMDB-51. All accuracies are averaged over three splits.

		DCE-101	HMDR-51			
Sal-4 Sol. BashXx1-01 3D 90.7		42.4	17.1	Method	Dim	UCF-101
87.7 59.1 3D 82.3	.,	84.4				
89.6 63.5 Two-szeam CNN [20] 2D 88.0 87.6 59.6 TDD [27] 2D 90.3		89.3	61.0	P3D [19]	3D	88.6
87.6 59.6 TDD [27] 2D 90.3				Two-stream I3D [3]	3D	98.0
81.0 39.0						

trained ResNet-18 clearly outperformed one trained from scratch. This result indicate that pretraining on Kinetics is effective on UCF-101 and HMDB-51. We can also see that effective on UCF-101 and HMDB-51. We can also see that the accuracies bestudely improved as the depth increased, similar to the results obtained on Kinetics. However, un-like the results on Kinetics, Resbelt-20 also improved the accuracies in HMDB-51. Because, as described above, the fine-tuning in this experiment was only performed to train corn²5, and the fully connected layer, the numbers of trained parameters were the same from Resbelt-50 in Resbelt-200. Therefore, the pretrained early layers, which may be a superior of the results of the results of the results of the major of the results of the results of the results of the results of the major of the results of the results of the results of the results of the major of the results of the results of the results of the results of the major of the results of the results of the results of the results of the major of the results of the results of the results of the results of the major of the results of the result

We shows the results of our comparison with state-of-the-art methods in Table 5. Here, we can see that ResNeXt-1101 (1997) and the shewed higher accuracies compared with C3D [23], P3D [19], two-stream CNN [20], and ITDD [27]. Further-more, we can also see that ResNeXt-101 (64f), which utilize

2D CNNs pretrained on ImageNet, even though the results of 3D CNNs trained on the UCF101 and HMDB51 datasets

3. Experimental configuration

In this study, in order to determine whether current video datasets have sufficient data for training of deep 3D CNNs, we conducted the three experiences described below using UCF-101 [21], HMDB-51 [17], AenviryNet [5], and Kinetics [16]. We first examined the training of relatively stallow 3D CNNs from serands or each video dataset. Aesthology 3D CNNs from serands or each video dataset. Aesthology 3D CNNs from serands or each video dataset. cording to previous works [9, 16], 3D CNNs trained on UCF-101, HMDB-51, and ActivityNet do not achieve high CCF-101, HSMD8-51, and ActivityNet do not dathere light accuracy whereones transien of kittens wat with the con-curacy whereones transien of kittens wat with the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the con-traction of the contraction of the contraction of the contraction of the contraction of the con-traction of the contraction of the Section 4.1 for details.

We then conducted a separate experiment to determine whether the Kinetics dataset could train deeper 3D CNNs. A main point of fish triul was to determine how deep? the datasets could train 3D CNNs. Therefore, we trained 3D seaSebses on Kinetics while varying the model depth from 18 to 200. If Kinetics can train very deep CNNs, such as ResNb-125, which adolested the best performance in as ResNet-152, which achieved the best perfortance in ResNets on ImageNet [10], we can be confident that they have sufficient data to train 3D CNNs. Therefore, the results of this experiment are expected to be very important for the future progress in action recognition and other video tasks. See Section 4.2 for details.

In the final experiment, we examined the fine-tuning or kinetics pretrained 3D CNNs on UCF-101 and HMDB-51 since pretraining on large-scale discisses is an effective way to achieve good performance levels on small disasses, we exto achieve good performance levels on small datasets, we ex-pect that the deep 3D ResNets pertained on Kinetics would perform well on relatively small UCF-101 and HMDB-51. This experiment examines whether the transfer visual rep-resentations by deep 3D CNNs from one domain to another domain works effectively. See Section 4.3 for details.

with 4D contounnous used in into study. Reserve, wines is one of the most successful architectures in image classi-fication, provides shortcut connections that allow a signal to bypass one layer and move to the next layer in the se

- miniplest networks for value action recognition. In Proceed-ings of the IEEE Conference on Companier Vision and Pattern Recognition (CVPR), 2017. 3, 8 C. Feichtenhofer, A. Piuz, and A. Zisserman. Convolutional two-stream network finites for video action recognition. In

- Josphenec on Computer 1996, 24, 5, 7

 G. Huang, Z. Liu, L. van der Manten, and K. Q. Weinberger

 Handrad convolutional networks. In Proceeding.
- Densyl contexted convolutional networks. In Proceedings of the IEEE Conference or Computer Vision and Fattern Recognition (CVPR), pages 4700-4108, 2017, 2, 4, 5, 7

 [13] S. Heft and C. Stragell, Baths of committation: Accelerating deep network training by reducing internal convariate shifts. In Proceedings of the International Conference on Many of the Internation

A basic RosNets block consists of two completional la tional layer is followed by batch nalization and a ReLU. A shortcut pass conn

tional lawer is followed by batch normalization and a ReLU, whereas each batch norm activation ResNet is followed by the ReLU and a co

evaluated.

The WRN architecture is the same as the ResNet (bost-tieneck), but there are differences in the number of feature maps for each corroditional layer. WRN increases the number of Sature maps rather than the number of Juyers. Such wide architectures are efficient in parallel computing using GPUs [31]. In this souty, we evaluate the WRN-50 using a widering factor of two.

3.3. Implementation

acilitates optimization in the training and reduces over-liting [11]. In this study, pre-activation ResNet-200 was

widening factor of two.

RosNeXt introduces cardinality, which is a different di-numerision from deeper and wider. Unlike the original bot-tleneck block, the ResNeXt block introduces group convo-lutions, which divide the feature maps into small groups. Cardinality refers to the number of middle convolutional layer groups in the bottleneck block. In their study, Xie et

layer groups in the bottleneck block. In their study, Xive of a showed that increasing the carallating of De-enhorizones is more effective than using some of the carbon study, we carbon from the continuity of the study, we carbon from early layers to late the study, we carbon from early layers to the Denvelocation than the contentiation from early layers to the players by the see of a concensation that is different from the ResNets summation. This concentration tens connectives all pages with the players of the see of a concensation that is different from the ResNets summation. This concentration tens connectives all page the pre-activation tend in pre-activation file. Next. In their layers of the pre-activation tend in pre-activation file. Next. In their layers of the pre-activation tend in pre-activation file. Next. In their layers of the pre-activation tend in pre-activation file. Next. The pre-activation tend in pre-activation file. Next. The pre-activation tend in pre-activation file. The pre-activation file and the pre-activation file.

Training. We use stochastic gradient descent with morn

um to trait the networks and randowly generate training samples from videos in training data in order to perform data augmentation. First, we select a temporal position in a video by uniform amplies in order to agreeate a train-saction of the contract of the contract of the contract selected temporal position. If the video is otherer than 16 of firms, show we hope its many times as accourage, beautiful entraction of the contract of the contract of the contract of the corner. In addition to the quality position from the 4 corners or a uputal scale of the sample in order to perform multi-scale copping. The precoducts used is the same at 13th. The that the sample width and height are the same as the shor und the single-would and engin are see some as the Snort-side length of the frame, and scale O.5 means that the sample is half the size of the sites riske length. The sample aspect ratio is I and the sample is segotis-temporally corpoped at the positions, scale, and aspect ratio. We spitially sestie the sample at 112×112 pitests. The size of each sample is 3 channels × 16 frames × 112 pitests, and each sample is horizontally flipped with 50% peochability. We also perform mean subteaction, which means that we subtract the nean values of ActivityNet from the sample for each colo channel. All generated samples retain the same class labels

In our training, we use cross-entropy losses and back m out thating, we use 'Cross-eatings' assess data teast-propagate their gradients. The training parameters include a weight decay of 0.001 and 0.0 for mentionism. When cratted 1, and observe which the contraction of the cratted 1, and observe which the contractions to the cratter of 0.001, and seep as weight decay of lee 5. Recognition. We adopt the bilding window manner to gra-responding to the contraction of the second of the contraction of t

Recognition. We adopt the sliding window anamer to gest-erate input (lap. (i.e., each whole is 1981 into non-overlapped 16-frame clips), and recognize actions in videos using the tazined networks. Each (lip is spatially) cropped around a center position at scale. I. We then input each clip into the networks and estimate the clip (to see, soccess, which are aver-aged over all the clips of the video. The class that has the maximum score indicates the recognized class label.)

DenseNet-[121, 201] DenseNet

- [17] H. Kuchne, H. Ruang, E. Garrote, T. Poggio, and T. Serre

04 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1

Figure 3: Block of each architecture. We represent costs, x2, F as the kernel size, and the number of feature maps of the co

 $A.3 \times 3 \times 3$ max-pooting layer (stride 2) is also located before conv2.x of all networks for down-sampling. In addition, conv1 spatially down-samples inputs with a spatial stride of two. C of the fully-connected layer is the number of classes.

| 10.8-x1 | 10.8-x1 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 2.

64 [6,6] 128 [12,12] 256 [24,48]

64 [2,3] 128 [2,4] 256 [2,6] 512 [2,3]

(16, 32)

can inclinate the staming of very deep networks. Citing previous studies that examined only limited 3D Resched architectures [9, 24], we examine not only deeper architectures [9, 24], we examine not only deeper architectures. By activities, we expire the following architectures: Resched thusk: and bottleneck blocks; [01], gre-activation Resched [11], while Resched (WRN) [11], Resched [19], and Demoched [12]. The architectures are summarized in Figure 3 and Table 1. In the following peragraphie, we will briefly introduce each in the following peragraphie, we will briefly introduce each.

- dataset of 101 human action classes, from videos in the wild. CRCV-TR, 12-01, 2012, 12, 23, 35

 C. Szegedy, W. Liu, Y. Jiu, P. Sermanet, S. Reed, D. Anguelov, D. Erfan, V. Vanhoueke, and A. Rabbosskin. Going deeper with currentitations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9, 2015, 3
- [25] D. Limit, Li. Seesaway and Control States with 3D consolutional na-learing spatiateraporal features with 3D consolutional na-works. In Proceedings of the International Conference on Camputer Vision (ICCV), pages 4480–4497, 2015. 2, 8
 [24] D. Tran, J. Ray, Z. Shen, S. Chang, and M. Polart. Convent architecture search for spatiotemporal feature learning. arXiv: 2017.01.2017, 2017. 2, 1, 4

- d action classes and instances.

 We resized the videos to heights of 240 pixels without bancing their associated and then stored them.

 Since the above mentioned experiment

4. Results and discussion

- 4.1. Analyses of training on each dataset We begin by training RosNot-18 on each dataset. Ac-cording to previous works 19, 163, 3D CNNs trained on UCF-101, HMD8-51, and ActivityNet do not achieve high accuracy whereas ones trained on Kinetics work well. We inted to reproduce such results in this experiment. In this process, we used split 1 of UCF-101 and HMD8-51, and

- - ment of accuracies compared with ResNet-18, which he previously examined architecture [9, 24]. In con-
- perform ResNeXI-101 even though ResNeXI-101 is a deeper architecture than ED. One of the reasons for this is the size differences of the network agrees. Specifically, the size of ED is $3 \times 64 \times 234 \times 234$, whereas that of ResNeXI-101 is $3 \times 16 \times 112 \times 112$. Thus, ED is 64 times larger than ResNeXI-101. To confirm the accuracies when using larger inputs, we also evaluated the ResNeXI-101 that used $3.64 \times 112 \times 112$ regular, which are the larger attaillib state. in our environment (NVIDIA TITAN X × 8). We can see
- Finally, in this section we confirm the perform UCF-101 and HMDB-51. The results of this experiment are important for determining website the 3D CNNs's are effective for other datasets. It should be noted that, in this cost, and the fully connected layer because it achieved the best performance, the area of the fully connected layer because it achieved the best performance during the preliminary experiments.

 Table 4 shows the accuracies of Kinetics pertained 3I. CNNs, as well as ResNes-18 trainfer from scratch, in UCF-CNNs, as well as ResNes-18 trainfer from scratch, in UCF-CNNs, as well as ResNes-18 trainfer from scratch, in UCF-CNNs, as well as ResNes-18 trainfer from scratch.
- school od. In Selection, for permanel couly inject, visited manages. These revolutions that for former representations of Roches 200 would be entailed for HIGHES 51 cent indicates the result of the selection of Roches 200 would be entailed for HIGHES 51 cent indicates the least of the result of

that two-stream I3D [3], which utilizes simple two-st

5. Conclusion

In the sindy, we assumed the architectures of various CNNs with spatio-empreed 3D consolitated laterate on CNNs with spatio-empreed 3D consolitated laterates on CNNs with spatio-empreed 3D consolitates confirming for engineerings, the filtering conclusions could be refused on the spatial of spatial concerning for the CNNs of the CNNs of

geNet experienced significant progress in various tasks such as object detection, semantic segmentation, and image cap-tioning. It is felt that, similar to these, 3D CNNs and Kinet-ics have the potential to contribute to significant progress in acs have the potential to contribute to significant progress in fields related to various video tasks such as action detection, video summarization, and optical flow estimation. In our future work, we will investigate transfer learning not only for action recomition but also for other such backs.

- Reterences
 11. S. Arto-Bi-Islija, N. Kerhart, J. Lee, P. Nasor, G. Toderict, B. Virndargan, and S. Viguneravinstan. You Thor-Sid: A large-scale video conditionals nethration. Activ. preprint, arXiv:1609.08615, 2016. 3
 21. Current and A. Zimerman. Quo vadio, action recognition? A new model and the Kinetics dationt. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (VPPI), pages 4724–4735, 2017. 2, 3
 [3] S. Current and A. Zimerman. Quo valid. action encognition?
- mon's how model and the Scottics affects and Experimental

 1. Deep, W. Deep, S. Scotte, L. J. B., K. L. and I. Feifer,
 Insignifies, A. Lago Scient Estimation lamp Database. In

 John S. L. Barris, S. Scotte, L. J. B., K. L. and I. Feifer,
 Insignifies, A. Lago Scient Estimation lamp Database. In

 John S. Congolius C. Scott, C. Scott, C. Scott, C. Scott,
 J. B. G. Datas C. Sch Bellense, Virge Feneral and J. C.

 1. B. G. Datas C. Sch Bellense, Virge Feneral and J. C.

 1. B. G. Datas C. Sch Bellense, Virge Feneral and J. C.

 Johnson Secophism (Exp. Prince-deep of the MEEI ConJourney on Company Training and Prince-Scotting of MEEI ConJourney on Company Training and Prince
 Journey on Co

- Emotion Recognition, 2017, 2, 3, 4, 6

 [10] K. Hu, X. Zhang, S. Rot, and J. Sun. Deep residual learning
 for irange recognition. In Proceedings of the IEEE Conference
 on Company Vision and Patrice Recognition (CVPR), pages
 770–778, 2016, 1, 2, 4, 5

 III. K. Bu, X. Zhang, S. Ben, and J. Sun. Identity inapplings
 in deep residual networks. In Proceedings of the European
 in Computer Vision (ECCVI), pages 600–645,
 2016, 2, 45. Computer Vision (ECCVI), pages 600–645,
 2016, 2, 45.

- 79, 2013. 6
 [27] L. Wang, Y. Qiao, and X. Tang. Action recognition will

その後は重要そうな部分に目を通す

- 手法, 結果, 結論など図表を中心に情報量の多い部分から目を通す

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

Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh National Institute of Advanced Industrial Science and Technology (AIST) Tsukuba, Ibaraki, Japan

creation of successful vision-based algorithms. In addition to such large-scale datasets, a large number of algorithms, such as residual learning [10], have been used to improve

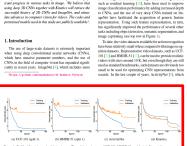
Figure 4 shows the training and validation loss lesNet-18 on each dataset. As can be seen in the figure alidation losses on UCF-101, HMDB-51, and Activity

quickly converged to high values and were clearly highe

specially concepted to high values and were clearly higher cone that specifing promised with the translation of the other cone that specifing promised where the training on those three datasets. In addition to those looses, we confirmed pre-chip control to the control of the control of the control of the specific property of the control of the specific property of the control of the control of the control of the excursion in property of the control of the control of the excursion in property of the control of the control of the excursion in property of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the control of the excursion in the

ement of accuracies compared with ResNet-18, which the previously examined architecture [9, 24]. In con-

We examine the architectures of various 3D CNNs from rel-atively shallow to very deep ones on current video datassets. Based on the results of those experiments, the following conclassions could be obtained: (1) ResNet-18 training r in significant overfitting for UCF-101_HMDR-51_ and Actiologible has not for Econics. (a) The Econics datume the indigition takes for instances of open 20 CONs, and enables moving of up to 12 SecOnic Inserts, received by substrained and up to 12 SecOnic Inserts, received by substrained and up to 12 SecOnic Inserts, received by substrained and up to 12 SecOnic Inserts are (10 SecIns) are set (10 SecIns) are substrained up to 12 SecOnic Inserts are (10 SecIns) are substrained and the SecOnic Inserts (10 SecIns) are substrained and the SecOnic Inserts (10 SecIns) and the SecInserts (10 SecIns) and the SecInsert (10 SecIns) and the SecInserts (10 SecInsert) and the SecInserts (10 SecInserts) and the SecI witeNet but not for Kinetics. (ii) The Kinetics dataset has



are portided in this chance.

Activityble (1/2) provides samples from 200 human action classes with an overage of 137 unrimmed video per class and 141 activity issuances per video. Unlike the other dances, Activityble consists of unrimmed video, which include non-action frames. The nord video length is 3-90 hours, and the total number of action instances is 28,108. This chances is randomly upli nin them deflerent subsects training, validation, and returning. More specifically, 50% is used for validation, and 25% of the control of the contro

used for raining, 25% in used for validation, and 25% is, consistently as the contract of the

We began by training ResNet-18 on each dataset. Ac we ogan by training research to an each trainer. Ac-cording to previous works [9, 16], 3D CNS trained on UCF-101, HMDB-51, and ActivityNet do not achieve high accuracy whereas ones trained on Kinetics work well. We tried to reproduce such results in this experiment. In this process, we used split 1 of UCF-101 and HMDB-51, and cale of video datasets has begun to approach that of imag

dimensional (3D) convolutional kernels (3D CNNs) are re-cently more effective than CNNs with two-dimensional (2D) kernels [2]. From several years ago [14], 3D CNNs are explored to provide an effective tool for accurate action recognition. However, even the usage of well-organized models [23, 25] has failed to overcome the advantages of 2D-based CNNs that combine both stacked flow and RGI 2D-based CNNs that combine both stacked flow and fields images [50]. The princip reasons for fish faith that been the images [50]. The princip reasons for fish faither that ear man be a similar and the composition of the c

However, can 3D CNNs retrace the successful history of 2D CNNs and ImageNet? More specifically, can the use of 3D CNNs trained on Kinetics produces significant progress in action recognition and other version used: The behavior in Figure 1.7 for Action with Pigure 1.7 for Action with Pigure 1.8 for the Figure 2.6 for Action with Pigure 2.6 for Action 1.7 for CNN should be a large-scale as interest particular to the Pigure 2.6 for Action 1.7 for Action 1.7

the training accomplished by 2D ResNets on ImageNet, a

ties of future progress in action recognition and

training of very deep 3D CNNs from scratch for action recog-nition. Previous studies showed deeper 2D CNNs trained on ImageNet achieved better performance [10]. However, it is not trivial to show deeper 3D CNNs are better based on the previous studies because the data-scale of image datasets differs from that of video ones. The results of this study, which indicate deeper 3D CNNs are more effective, can be expected to facilitate further progress in computer vision for

2.1. Video Datasets

The HMDB-51 [17] and UCF-101 [21] datasets are cur rently the most successful in the field of action recogni tion. These datasets gained significant popularity in the early years of the field, and are still used as popular benchmarks. they are simply not large enough for training deep CNNs

introduced, larger video datasets were produced. These include ActivityNet [5], which contains 849 hours of video,

cluding 25,000 action instances, Activity Next also provides now additional tasks, such as untrimmed classification and exection, but the number of action instances is still on the dress of the set of the contrast. This year (2017), in an effort to roate a successful pretrained model, Kay of al. elosaed the institis claused 161. The Kinetics classes includes more an 900,000 trimmed videos covering 400 categories. In their to determine whether it can train deper 30 CNNs, or performed a number of experiments using those recent ets, as well as the LICE-101 and HMDB-51 dataset

tatasets, as well as the UCF-101 and IMOID-51 datasets. Other high datasets such as Spect-104 [15] and Gralibet-8-68 [11] have been perposed. Although these altrabases are larger from Knietts, their antontions are translations are larger from Knietts, their antontions are supported to the such as the support of the such as the support of the support in from discussing these datasets in this study.

One of the popular approaches to CNN-based action recognition is the use of two-stream CNNs with 2D convolutional kernels, in their study, Simonyan et al. proposed a method that uses RGB and stacked optical flow formes as appearance and motion information, respectively [20], and showed that combining the two-streams has the ability to improve action recognition accuracy. Since that study, numerous methods based on the two-stream CNNs

study, uninerous methods based on the two-steam CNNs have been proposed to improve action recognition performance (6, 7, 8, 27, 28, 29). Utilité the showmentioned approaches, we focused on CNNs with 3D correductional learnets, which have recently began to comprehen 2D CNNs from an 2D CNNs from the performance video demonst. These 3D CNNs are intutively effective because such 3D convolution can be used to directly extract because such 3D convolution can be used to directly extract hat using optical flows as inputs to 3D CNNs resulted in a ligher level of performance than can be obtained from RGB inputs, but that the best performance could be achieved by combining RGB and optical flows. Meanwhile, Kay et al. showed that the results of 3D CNNs trained from scratch on

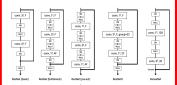
their Kisetics dataset were comparable with the results of 2D CNNs perearized on ImagNet, even though the results of 3D CNNs trained on the UCP101 and BMDB31 datasets were inferior to the 2D CNNs results. In still another study, Carstrian et al. proposed inception [25] based 3D CNNs, which they referred to as EID, and achieved state-of-the-art performance [21]. There recently, usen works introduced ResNet architectures into 3D CNNs [9, 24], though they examined unly reliefsively Sulfation ones.

3. Experimental configuration

In this study, in order to determine whether current video datasets have sufficient data for training of deep 3D CNNs, we conducted the three experiments described below using UCF-101 [21], HMDB-51 [17], ActivityNet [5], and Kinetics [16]. We first examined the training of relativity shallow 3D CNNs from strank on each video dianet. Section 4.1 for details

we then conducted a separate experiment to determine whether the Kinetics dataset could train deeper 30 CNNs. A main point of this trial was to destrainte how deepy the datasets could train 30 CNNs. Therefore, we trained 30 ResiNess on Kinetics white varying the model depth from 18 to 200. If Kinetics can train very deep CNNs, such as ResNet-152, which achieved the best performance in ResNets on ImageNet [10], we can be confident that the have sufficient data to train 3D CNNs. Therefore, the resul

to achieve good performance levels on small datasets, we ex-pect that the deep 3D ResNets pertained on Kinetics would perform well on relatively small UCF-101 and HMDB-51. This experiment examines whether the transfer visual rep-resentations by deep 3D CNNs from one domain to another domain works effectively. See Section 4.3 for details.



Model	Block	conv1	00	rv2_x	0	onv3_x		oev4_x	con	n5_x	
			F	N	F	N	F	N	F	N	
ResNet- (18, 34)	Basic		64	{2, 3}	128	{2,4}	256	(2, 6)	512	{2, 3}	
ResNet-{50, 101, 152, 200}	Bottleneck	ぎさん	64	3	128	{4, 4, 8, 24}	256	{6,23, 36,36}	512	3	pool.
Pre-act ResNet-200	Pre-act	conv, 7×7×7, c temporal stride spatial stride	64	3	128	24	256	36	512	3	global average pool. C-d fally-connected.
WRN-50	Bottleneck	temporal spatial	128	3	256	4	512	6	1024	3	24
ResNeXt-101	ResNeXt	§ E #	128	3	256	24	512	36	1024	3	85
DenseNet- [121, 201]	DenseNet		64	{6,6}	128	(12, 12)	256	{24, 48}	{512, 896}	{16, 32}	W 0

pievono studies that elemente only influent 2 in Rescue al-ritherature [9, 24], we examine not only desper architectures but also some extended versions of RenNer. Re particular, we explore the following architectures: ResNet (basis: and bottleneck blocks) [10], pre-activated ResNet [11], wide ResNet (WRN) [31], ResNext [30], and Demoshet [12]. The architectures are summarized in Figure 3 and Table 1. In the following peragraphs, we will briefly strondere each

3.3. Implementation

A basic RosNots block consists of two completional lan tional layer is followed by batch n nalization and a ReLU. A shortcut pass conn

multiplier and all ReLU. A shortest pure conversely the type of the book in the large part belief to the ReLU in the book. Reaches 18 and 34 adopt the base book. We not selected the Reaches 18 and 34 adopt the base book. We not selected the Reaches 18 and 34 adopt the base book. We not selected the relationship of the Reaches 18 and 18

a ReLU, whereas each batch normalization of the pr activation ResNet is followed by the ReLU and a co

GPUs [31]. In this study, we evaluate the WRN-50 using a widening factor of two.

mension from deeper and wider. Unlike the original bot-tleneck block, the ResNeXt block introduces group convo-Cardinality refers to the number of middle conv ayer groups in the bottleneck block. In their study, Xie et

of 32. Dense/Net makes connections from early layers to later layers by the use of a concatenation that is different from the RenNets summation. This concatenation connects each layer densely in a feed-forward fashion. Dense/Nets also adopt the pre-activation used in pre-activation lessNets. In their Study, Huang et al. showed that it achieves better accuracy

Training. We use stochastic gradient descent with mome

num to train the networks and randomly spoarnes training samples from visions in training data in order to perform data augmentation. First, we select a temporal position in a video by uniform sampling in order to guestras attain-ing sample. A 16-frame clip is then guestrast attain-ing sample. A 16-frame clip is then guestrast around the selected emporal position. If the video is shorter than 16-frames, then we loop it as many times as necessary. Next, we randomly select a spatial position from the 4 corners or the center. In addition to the spatial position, we also select a spatial scate of the sample in order to perform multi-scale. cropping. The procedure used is the same as 1281. The channel. All generated samples retain the same class labels

m out thating, we use 'Cross-eatings' assess data teast-propagate their gradients. The training parameters include a weight decay of 0.001 and 0.0 for mentionism. When cratted 1, and observe which the contraction of the cratted 1, and observe which the contractions to the cratter of 0.001, and seep as weight decay of lee 5. Recognition. We adopt the bilding window manner to gra-responding to the contraction of the second of the contraction of t

similar to those for 2D ResNets on ImageNet [11]. Mo

Comparisons with other architectures are shown in Ti ble 2. Here, it can be seen that the accuracies of pre-activation ResNet-200 are slightly low when compared with the standard ResNet-200 though the et al. reported that it pre-activation reduces overfitting and improves 2D ResNet 200 no ImageNet [11]. We can also see that the WRN-3 continued to the compared with the ResNet

compared with CSD with batch normalization [16], which is 10-layer network, as well as CNN-LSTM and two-stream CNN [16]. This result also indicates the effectiveness of edger 3D network strained on Kinetics. In contrast, RGB-DD trained on Kinetics from seraith [3] were found to outperform RenNix1-101 event though ResNix1-101 is a deeper architecture than 13D. One of the reasons for this is the size of the reasons for this production of the reasons for this in the size of the reasons for the reasons for this in the size of the reasons for the reasons fo unreconces or the network inputs. Specifically, the size of 18D is $3\times64\times224\times224$, whereas that of ResNeXt-101 is $3\times61\times112\times112$. Thus, 18D is 64 times larger than ResNeXt-101. To confirm the accuracies when using larger inputs, we also evaluated the ResNeXt-101 that used $3\times64\times112\times112$ inputs, which are the largest available sizes

Method	Top-1	Top-5	Averag
ResNet-18	54.2	78.1	66.1
ResNet-34	60.1	81.9	71.0
ResNet-50	61.3	83.1	72.2
ResNet-101	62.8	83.9	73.3
ResNet-152	63.0	84.4	73.7
ResNet-200	63.1	84.4	73.7
ResNet-200 (pre-act)	63.0	83.7	73.4
Wide ResNet-50	64.1	85.3	74.7
ResNeXt-101	65.1	85.7	75.4
DenseNet-121	59.7	81.9	70.8
DenseNet-201	61.3	83.3	72.3

Method	Top-1	Top-5	Average
ResNeXt-101			74.5
ResNeXt-101 (64f)	-	-	78.4
CNN+LSTM [16]	57.0	79.0	68.0
Two-stream CNN [16]	61.0	81.3	71.2
C3D w/ BN [16]	56.1	79.5	67.8
RGB-I3D [3]	68.4	88.0	78.2
Thron advances (SE) (C)	21.6	90.0	80.8

tures with ResNeXt-101 make further improvements base

4.3. Analyses of fine-tuning Finally, in this section we confirm the performance of UCF-101 and HMDB-51. The results of this experiment are important for determining website the 3D CNNs's are effective for other datasets. It should be noted that, in this cost, and the fully connected layer because it achieved the best performance, the area of the fully connected layer because it achieved the best performance during the preliminary experiments.

Table 4 shows the accuracies of Kinetics pertained 3I. CNNs, as well as ResNes-18 trainfer from scratch, in UCF-CNNs, as well as ResNes-18 trainfer from scratch, in UCF-CNNs, as well as ResNes-18 trainfer from scratch, in UCF-CNNs, as well as ResNes-18 trainfer from scratch.

ResNet-18 (scratch)	42.4	17.1
ResNet-18	84.4	56.4
ResNet-34	87.7	59.1
ResNet-50	89.3	61.0
ResNet-101	88.9	61.7
ResNet-152	89.6	62.4
ResNet-200	89.6	63.5
DenseNet-121	87.6	59.0
ResNeXt-101	50.7	63.2

ResNet-18 clearly outperformed one trained from This result indicate that pretraining on Kinetics is e on UCF-101 and HMDB-51. We can also see tha

88.0 90.3

	17.1	Method
•	17.1	ResNeXt-101
	56.4	ResNeXt-101 (64f)
	59.1 61.0	C3D [23]
,	61.7	P3D [19]
5	62.4	Two-stream I3D [3]
5	63.5	Two-stream CNN [20]
5	59.6	TDD [27]
7	63.8	ST Multiplier Net [7]

acs have the potential to contribute to significant progress in fields related to various video tasks such as action detection, video summarization, and optical flow estimation. In our future work, we will investigate transfer learning not only for action recognition but also for other such tasks.

Reiterenzee, B. Variaterjas, N. Kerhart, J. Lee, P. Natoer, G. Toderici, B. Variaterjas, and S. Vigopramarinthas, YouThen-SM: A lago-scale video conditionals nethratine. dockin preprint, arXiv:1609.08615, 2016. 3
2.1 J. Curretta and A. Zionerman. Quo vadis, action recognition? A new model and the Kinteis dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (VPPR), pages 4724–4735, 2017. 2, 3
[3] A. Curretta and A. Zionerman. Quo vadis, action ecognition?

AND CONTROLLED AND CO

Denoity connected convolutional networks. In Proceedings of the IEEE Configurous or Comparer Vision and Flattom Economism (CVPR), pages 4700–4708, 2017, 2, 4, 5, 7 [13] S. bife and C. Sergods. Buth commitations: Accelerating deep network training by reducing internal convariant shiftle. Leveling, pages 485–485, 2015, 4 [14] S. R. W. Xii, M. Tang, and K. Yu. 2D correlational neural

dataset of 101 human action classes from videos in the wild. CRCVTR-12-01, 2012. 1, 2, 3, 5
2] C. Srepedy, W. Liu, Y. Jia, P. Soemanet, S. Reed, D. Anguelov, D. Ethan, V. Vanhoucke, and A. Robinswich. Going deeper with correlations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9, 2015. 3

[24] D. 1881. Le overage and Estates with 3D convolutional na-works. In Proceedings of the International Conference on Computer Vision (ICCV), pages 4489–4497, 2015. 2, 8
[24] D. Tran, J. Ray, Z. Steu, S. Chang, and M. Palari. Convoct sechiecture search for spatiotemporal feature learning. arXiv

79, 2013. 6
127| L. Wang, Y. Qiao, and X. Tang. Action recognition will

ポイントは意外と限られる?

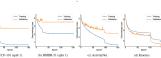
– In this study (paper/research), ∼ や Our contribution(s)∼ など序盤に現れる

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

Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh National Institute of Advanced Industrial Science and Technology (AIST) Tsukuba, Ibaraki, Japan

significant overfating for UCF-101, HMDB-51, and AciteNet but not for Kinetics. (ii) The Kinetics dataset ha trout/tvet but not for Knetics. (ii) the Knetics dutaset hus sufficient data for training of deep 3D CNNs, and enables training of up to 152 ResNets layers, interestingly similar to 2D ResNets on ImageNet. ResNeXi-101 achieved 78.4% average accuracy on the Kinetics test set. (iii) Kinetics pre-





ing, validation, and testing sets are about

We began by training ResNet-18 on each dataset

Figure 4 shows the training and validation loss sNet-18 on each dataset. As can be seen in the figure lidation losses on UCF-101, HMDB-51, and Activit quickly converged to high values and were clearly highe

gnition. However, even the usage of well-organized lels [23, 25] has failed to overcome the advantages of 3D convolution that can be engaged by the Kinetics dataset

by 2. Here, it can be seen that the accuracies of pea-activation ResNet-200 are slightly low when compared with the standard ResNet-200 though He et al. reported that the

Table 3 shows the results of the Kinetics test set used CNN [10]. This result also indicates the effectiveness of deeper 3D networks trained on Kinetics. In contrast, RGB-I3D trained on Kinetics from scratch [3] were found to out-perform ResNeXt-101 even though ResNeXt-101 is a deeper architecture than 13D. One of the reasons for this is the size differences of the network inputs. Specifically, the size (I3D is $3 \times 64 \times 224 \times 224$, whereas that of ResNeXt-DI is $3 \times 16 \times 112 \times 112$. Thus, I3D is 64 times larger

RGB-I3D [3]

including 23,000 action stances. Activity Net also provides some additional tasks, such as untimmed classification and detection, but the unknown of action instances is still on the other other other other control of the control of the other o asets as well as the LICE-101 and HMDR-51 dataset

their Kiterics dataset were comparable with the results of 2D CNNs premiated on ImageNet, even though the results of 3D CNNs trained on the UCF101 and HMDB31 datasets were interior to the 2D CNNs results in 10d another study. Carriera et al. proposed inception [22] based 3D CNNs, which they refore to a tell 7m and advised vature-of-the-art performance [2]. More recently, some works introduced ResNet architectures into 3D CNNs [9, 24], though they cantinoid only reformly shallow once

A basic RosNets block consists of two c nalization and a ReLU. A shortcut pass conne

multation and a BeLLI. A shortest pose connects the type of the book to the but per between the last Ed. In the book. ResNet-18 and 34 adapt the book theck. We me learning the book to the but per between the book to the book to book type A in [10] to sootd increasing the number of parameters of these relatively shallow networks. A BeSNets Distincted book consists of the convolu-tional upon. The levned store of the first and thad convo-tant and the store of the store of the store of the store and 3 x 3 x 3. The Store of the store of the store of that of the boule book. He solved the store of for those the challenges will be designed to the store of for those

Training. We use stochastic gradient descent with a samples from videos in training data in order to perfort data augmentation. First, we select a temporal position i

In this study, we examine various 3D CNN architectures from relatively shallow to very deep ones using the Kinetics and other popular video datasets (UCF-101, HMDB-51, and ActivityNet) in order to provide us insights for answering the above question. The 3D CNN architectures tested in this study are based on residual networks (ResNets) [10] and their extended versions [11, 12, 30, 31] because they have simple and effective structures. Accordingly, using those datasets, we performed several experiments aimed at training and testing those architectures from scratch, as well

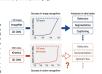
ポイントは意外と限られる?

- 関連研究の最後に従来との比較・差分が書かれる(ことが多い)

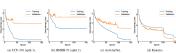
Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?

National Institute of Advanced Industrial Science and Technology (AIST) Tsukuba, Ibaraki, Japan

significant overfating for UCF-101, HMDB-51, and Ac-vityNet but not for Kinetics. (ii) The Kinetics dataset has trout/tvet but not for Knetics. (ii) the Knetics dutaset hus sufficient data for training of deep 3D CNNs, and enables training of up to 152 ResNets layers, interestingly similar to 2D ResNets on ImageNet. ResNeXi-101 achieved 78.4% average accuracy on the Kinetics test set. (iii) Kinetics pre-



creation of successful vision-based algorithms. In addition to such large-scale datasets, a large number of algorithms such as residual learning [10], have been used to improve

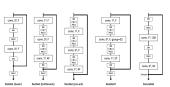


ie began by training ResNet-18 on each dataset

refrain from discussing these datasets in this study.

including 28,000 action instances. Activity Net also provides some additional tacks, such as untrimmed classification and detection, but the number of action instances is still on the order of tens of thousands. This year (2017), in an effort to create a successful pertrained model, Ray or al. released the Kinetics dataset [16]. The Kinetics dataset includes more

Other huge datasets such as Sports-1M [15] and YouTube-8M [1] have been proposed. Although these databases are larger than Kinetics, their annotations are slightly noisy and only video-level labels have been assigned. (In other words, they include frames that do not relate to target actions.) Such noise and the presence of unrelated frames have the potential to prevent these models from providing good training. In addition, with file sizes in excess of 10 TB, their scales are simply too large to allow them to be utilized easily. Because of these issues, we will



del	Block conv1		conv2_x			con/3_x		conv4_x		com5_x	
	DIAL.		F	N	F	N	F	N	F	N	
sNes- 8,341	Blss		64	{2, 3}	128	{2,4}	256	{2, 6}	512	{2, 3}	
rsNet-{50, 1, 152, 200}	Bottleneck	3 - 1	64	3	128	{4, 4, 8, 24}	256	{6,23, 36,36}	512	3	pool.
e-act rsNet-200	Pre-act	7×7,0 4 stride stride	64	7	128	24	256	36	512	3	global average pool. C-d fully-connected,
RN-50	Bottleneck	nporal spatial	128	3	256	•	512	6	1024	3	100
sNeXt-101	ResNeXt	spatial	128	3	256	24	412	36	1024	3	8 P
mseNet-	DenseNet	0 -	64	(6, 6)	128	(12, 12)	256	[24-48]	{512,	{16, 32}	90 C

3.3. Implementation

A basic RosNets block consists of two c ers, and each convolutional layer is followed by batch i malization and a ReLU. A shortcut pass connects the to he block to the layer just before the last Rel.U in the block ResNet-18 and 34 adopt the basic blocks. We use identit

Training. We use stochastic gradient descent with

その分野に長けていれば...

- 引用されている論文や場所を見ればなんとなく内容が分かる?

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

Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh National Institute of Advanced Industrial Science and Technology (AIST) Tsukuba, Ibaraki, Japan (kensho.hara, hirokatsu.kataoka, yu.satou)@aist.go.jp

We examine the architectures of various 3D CNNs from rel-atively shallow to very deep ones on current video datassets. Based on the results of those experiments, the following conclassions could be obtained: (1) ResNet-18 training r in significant overfitting for UCF-101, HMDB-51, and Actiologible has not for Econics. (a) The Econics datume the indigition takes for instances of open 20 CONs, and enables moving of up to 12 SecOnic Inserts, received by substrained and up to 12 SecOnic Inserts, received by substrained and up to 12 SecOnic Inserts, received by substrained and up to 12 SecOnic Inserts are (10 SecIns) are set (10 SecIns) are substrained up to 12 SecOnic Inserts are (10 SecIns) are substrained and the SecOnic Inserts (10 SecIns) are substrained and the SecOnic Inserts (10 SecIns) and the SecInserts (10 SecIns) and the SecInsert (10 SecIns) and the SecInserts (10 SecInsert) and the SecInserts (10 SecInserts) and the SecI witeNet but not for Kinetics. (ii) The Kinetics dataset has



creation of successful vision-based algorithms. In addition to such large-scale datasets, a large number of algorithms, such as residual learning [10], have been used to improve

such as residual latering [19], how been used to improve many classications professionally assignment by admit gracewood optimization and continuous contractions of the contraction of the contraction of generic facilities of the acquestion of generic facilities of the acquestion of generic facilities (as in the contraction), in turn, respectives, the contraction of the co

Figure 4 shows the training and validation losses esNet-18 on each dataset. As can be seen in the figure, didation losses on UCF-101, HMDB-51, and Activity

quickly converged to high values and were clearly higher

specially concepted to high values and were clearly higher cone that specifing promised with the translation of the other cone that specifing promised where the training on those three datasets. In addition to those looses, we confirmed pre-chip control to the control of the control of the control of the specific property of the control of the specific property of the control of the control of the control of the excursion in property of the control of the control of the excursion in property of the control of the control of the excursion in property of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the excursion in the control of the control of the control of the control of the excursion in the

Here, we will show ResNets accuracies changes based on model depths. Figure 2 shows the averaged accuracies over top-1 and top-5 ones. We can see that, essentially, as the depth increased, accuracies improved, and that the im-provements continued until reaching the depth of 152. We can also see that deeper ResNet-152 achieved significant im-

ement of accuracies compared with ResNet-18, which the previously examined architecture [9, 24]. In con-

icale of video datasets has begun to approach that of imag

For action recognition, CNNs with spatio-temporal threadmensional (30) convolutional kernels (3D CNNs) are re-cently more effective than CNNs with two-dimensional (20) kernels [21]. From several years ago [14], 3D CNNs are explored to provide an effective tool for accurate action recognition. However, even the usage of well-orgalized models [23, 25] has failed to overcome the advantages of 2D-based CNNs that combine both stacked flow and RGB 2D-based CNNs that combine both stacked flow and RGB, magas [50]. The principal reasons for this faither has been the ranges [50]. The principal reasons for this faither has been the region of the principal reasons for the faither has been the first principal reasons and the state of the region of the reasons and the reasons and the reasons and the reasons and the reasons are the

However, can 3D CNNs retrace the successful history of 2D CNNs and ImageNet? More specifically, can the use of 3D CNNs trained on Kinetics produces significant progress in action recognition and other version used: The behavior in Figure 1.7 for Action with Pigure 1.7 for Action with Pigure 1.8 for the Figure 2.6 for Action with Pigure 2.6 for Action 1.7 for CNN should be a large-scale as interest particular to the Pigure 2.6 for Action 1.7 for Action 1.7

those datasets, we performed several experiments aimed at training and testing those architectures from scratch, as well



Figure 2: Averaged accuracies of 3D ResNets over top-1 and topon the Kinetics validation set. Accuracy levels improve as net depths increase. The improvements continued until reaching depth of 152. The accuracy of ResNet-200 is almost the sar that of ResNet-152. These results are similar to 2D ResNets on

as their fine-tuning. The results of those experiments (see Section 4 for details) show the Kinetics dataset can train 3D ResNet-152 from scratch to a level that is similar to the training accomplished by 2D ResNets on ImageNet, a shown in Figure 2. Based on those results, we will discus ies of future progress in action recognition and

training of very deep 3D CNNs from scratch for action recog-nition. Previous studies showed deeper 2D CNNs trained on ImageNet achieved better performance [10]. However, it is not trivial to show deeper 3D CNNs are better based on the previous studies because the data-scale of image datasets differs from that of video ones. The results of this study, which indicate deeper 3D CNNs are more effective, can be expected to facilitate further progress in computer vision for

2.1. Video Datasets

The HMDB-51 [17] and UCF-101 [21] datasets are cur-rently the most successful in the field of action recogni-tion. These datasets gained significant popularity in the early years of the field, and are still used as popular benchmarks. However, a recent consensus has emerged that indicates tha they are simply not large enough for training deep CNNs

A couple of years after the abovementioned datasets wen introduced, larger video datasets were produced. These include ActivityNet [5], which contains 849 hours of video, including 28,000 action instances. ActivityNet also provides some additional tacks, onto a merriment classification and detection, but the number of action instances is still on the other of tens of themsessal. This year (2017), in an effort or cente a successful pretrained model, Kay et al. ebased the Knetics datused [16]. The Knetics disease if includes more than 900,000 trimuned videos covering 400 categories. In order to determine whether it can train deeper 3D CNNs, we perfermed a number of experiments using these recent e performed a number of experiments using these re-atasets, as well as the UCF-101 and HMDB-51 dataset

durient, as well as the UCF-101 and HMDB-51 datasets. Other lang datasets such as Spect-14M [15] and Vorflies-5M [1] have been proposed. Although those UCF-102 and UCF-102 an

One of the popular approaches to CNN-based action necognition is the use of two-stream CNNs with 2D con-nectional kernels. In their study, Simonyan et al. proposed as method that uses Roll and stacked optical flow frames as appearance and motion information, respectively [20], and showed that combining the two-streams has the abtlity to improve action recognition accuracy. Since that study, numerous methods based on the two-stream CNNs

study, uninerous methods based on the two-steam CNNs have been proposed to improve action recognition performance (6, 7, 8, 27, 28, 29). Utilité the showmentioned approaches, we focused on CNNs with 3D correductional learnets, which have recently began to comprehen 2D CNNs from an 2D CNNs from the performance video demonst. These 3D CNNs are intutively effective because such 3D convolution can be used to directly extract because such 3D convolution can be used to directly extract spatio-temporal features from raw videos. For example, Ji et al. proposed applying 3D convolution to extract spatio-CNNs, which they referred to as C3D, using the Sportsthat using optical flows as inputs to 3D CNNs resulted in a higher level of performance than can be obtained from RGB inputs, but that the best performance could be achieved by combining RGB and optical flows. Meanwhile, Kay et al. showed that the results of 3D CNNs trained from scratch on

their Kisetics dataset were comparable with the results of 2D CNNs perearized on ImagNet, even though the results of 3D CNNs trained on the UCP101 and BMDB31 datasets were inferior to the 2D CNNs results. In still another study, Carstrian et al. proposed inception [25] based 3D CNNs, which they referred to as EID, and achieved state-of-the-art performance [21]. There recently, usen works introduced ResNet architectures into 3D CNNs [9, 24], though they examined unly reliefsively Sulfation ones.

3. Experimental configuration

In this study, in order to determine whether current video datasets have sufficient data for training of deep 3D CNNS, we conducted the three experiences described below us-ing UCF-101 [21], HMDB-51 [17], ActivityNet [5], and Kinetics [16]. We first examined the training of relatively studies 3D CNNS from scranch on early video danset. According to previous works [9, 16], 3D CNNs trained on UCF-101, HMDB-51, and ActivityNet do not achieve high CCT-101, HMMb-51, and Activitybed to not active high accuracy where ones transiend festiciates work well. We try to reproduce such results to accertain whether the datasets are sufficient data for deep 3D CNNS. Specifically, we used ResPert 18, which is the shallowest ResNet architec-ture, based on the assumption that if the ResNet 3E overfits when heig trained on a dataset, that diameter is too small to be used for training deep 3D CNNs from scranch. See Section 4.1 for details

we then conducted a separate experiment to determine whether the Kinetics dataset could train deeper 30 CNNs. A main point of this trial was to destrainte how deepy the datasets could train 30 CNNs. Therefore, we trained 30 ResiNess on Kinetics white varying the model depth from 18 to 200. If Kinetics can train very deep CNNs, such as ResNet-152, which achieved the best performance in ResNets on ImageNet [10], we can be confident that the have sufficient data to train 3D CNNs. Therefore, the resul

to achieve good performance levels on small datasets, we ex-pect that the deep 3D ResNets pertained on Kinetics would perform well on relatively small UCF-101 and HMDB-51. This experiment examines whether the transfer visual rep-resentations by deep 3D CNNs from one domain to another domain works effectively. See Section 4.3 for details.

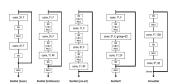


Figure 3: Block of each architecture. We represent cosy, x², F as the kernel size, and the number of feature maps of the co

Model	Block	conv1	00	conv2_x com		onv3_x		oes4_x	conv5_x		
			F	N	F	N	F	N	F	N	
ResNet- (18, 34)	Basic		64	{2, 3}	128	{2,4}	256	(2, 6)	512	{2, 3}	
ResNet-(50, 101, 152, 200)	Bottleneck	± = 4	64	3	128	{4, 4, 8, 24}	256	{6,23, 36,36}	512	3	pool,
Pre-act ResNet-200	Pre-act	7×7×7,0 oral stride tal stride	64	3	128	24	256	36	512	3	al average pool, fully-connected,
WRN-50	Bottleneck	my.7x mporal quital	128	3	256	4	512	6	1024	3	34
ResNeXt-101	ResNeXt	temporal spatial	128	3	256	24	512	36	1024	3	global C-d ful
DenseNet- [121, 201]	DenseNet	0 -	64	{6,6}	128	(12, 12)	256	{24, 48}	{512, 896}	{16, 32}	20.0

can facilitate the training of very deep networks. Unlike increases statistics and the examines of only deeper architectures but also some extended versions of ResNet. In particular, we explore the following architectures: ResNet (basis and bottleneck, blocks) [10], pre-activation ResNet [11], wide ResNet (WRN) [31], ResNeXt [30], and DemseNet [12]. The architectures are summarized in Figure 3 and Table 1. In the following paragraphs, we will briefly introduce each

3.3. Implementation

A basic RosNets block consists of two completional la tional layer is followed by batch n nalization and a ReLU. A shortcut pass conn

militation and a BeLLI. A shortest pass connects the top of the book on the large part before the last BeL III in the book. RenNet13 and 34 along the baxes blocks. We not estimate connections and zero pauling for the shortests of the basic connections and zero pauling for the shortests of the basic parameters of these fellinely shallow networks. A ResNets bottlessche block consists of these convolu-tional layers are 1 x 1 x 1, whereas those of the economi-tional layers are 1 x 1 x 1, whereas those of the econo-tical state of the best block. ResNet-SQ 10, 11, 22, and 20 adopt the bottlessch. We test leaving connections except for those whereas the state of the shortest of the state of the state of the state of the bottlessch. We test leaving connections except for those whereas the state of the sta

a ReLU, whereas each batch normalization of th activation ResNet is followed by the ReLU and a co acilitates optimization in the training and reduces over-liting [11]. In this study, pre-activation ResNet-200 was

GPUs [31]. In this study, we evaluate the WRN-50 using a widening factor of two.

ResNeXt introduces cardinality, which is a different di-mension from deeper and wider. Unlike the original bot-tleneck block, the ResNeXt block introduces group convo-Cardinality refers to the number of middle conv ayer groups in the bottleneck block. In their study, Xie et

DenseNet makes connections from early layers to later layers by the use of a concatenation that is different from the ResNets summation. This concatenation connects each layer densely in a feed-forward fashion. DenseNets also adopt the pre-activation used in pra-entration ResNets. In their study, Huang et al. showed that it achieves better accuracy

Training. We use stochastic gradient descent with mome

num to train the networks and randomly spoarnes training samples from visions in training data in order to perform data augmentation. First, we select a temporal position in a video by uniform sampling in order to guestras attain-ing sample. A 16-frame clip is then guestrast attain-ing sample. A 16-frame clip is then guestrast around the selected emporal position. If the video is shorter than 16-frames, then we loop it as many times as necessary. Next, we randomly select a spatial position from the 4 corners or the center. In addition to the spatial position, we also select a spatial scate of the sample in order to perform multi-scale. cropping. The procedure used is the same as 1281. The channel. All generated samples retain the same class labels

m out thating, we use 'Cross-eatings' assess data teast-propagate their gradients. The training parameters include a weight decay of 0.001 and 0.0 for mentionism. When cratted 1, and observe which the contraction of the cratted 1, and observe which the contractions to the cratter of 0.001, and seep as weight decay of lee 5. Recognition. We adopt the bilding window manner to gra-responding to the contraction of the second of the contraction of t

are 4: BenNet-18 training and validation looses. The validation looses on UCF-101, HMDB-51, and ActivityNet quickly on to values and were clearly higher than their corresponding training bosses. The validation looses on Kinetics were slightly be corresponding training looses, slightlicastly different learn those on the other datasets.

(a) DCF-101 (selfc1). (b) HMDR-51 (selfc1).

are portided in this chance.

Activityble (1/2) provides samples from 200 human action classes with an overage of 137 unrimmed video per class and 141 activity issuances per video. Unlike the other dances, Activityble consists of unrimmed video, which include non-action frames. The north video length is 3-30 loose, and the total number of action instances is 28,108. This chances is randomly upli nin them deflerent subsection with the contraction of the contraction

used for raining, 25% in used for validation, and 25% is, consistently as the contract of the

4.1. Analyses of training on each dataset

We began by training ResNet-18 on each dataset. Ac

we ogan by training rossore to on each trainer. Ac-cording to previous works [9, 16], 3D CNS trained on UCF-101, HMDB-51, and ActivityNet do not achieve high accuracy whereas ones trained on Kinetics work well. We tried to reproduce such results in this experiment. In this process, we used split 1 of UCF-101 and HMDB-51, and

ResNet-200 started to overfit. Interestingly, the results are similar to those for 2D ResNets on ImageNet [11]. More specifically, the accuracies of both 2D and 3D ResNets im-proved as the depth increased utility caching the depth of 152, and then the accuracies did not increase when increas-ing the depth of 200. These results indicate that the Kinet-ies densed has sufficient data to train 3D CNNs in a measure imilar to ImageNet.

Comparisons with other architectures are shown in Ta-ble 2. Here, it can be seen that the accuracies of per-activation ResNet-200 are slightly low when compared with the standard ResNet-200 though He et al. reported that the per-activation reduces overfitting and improves 2D ResNet-200 on ImagNet [11]. We can also see that the WRN-50 actives/d inject accuracies when compared with the BesNet-ted Processing and the process of the second process of the compared with the BesNet-ted Processing Processing and Processing Pro

Table 3 shows the results of the Kinetics test set used Table 3 shows the results of the Kinetics test set must to compare Backs(10), which acknowled the highest ac-curation, with the state-of-the-ort modulot. Here, it can be compared with CDD with both formatting [16], which is 10-layer network, as well as CDN-LSTM and two-steams CDN [16]. Then could also indicates the officiences of CDN [16]. Then could also indicates the officiences of DD trained on Kinetics from scrattle [1] were found to our perform Backs(2-CL) in the algorithm of the could be of perform Backs(2-CL) in the algorithm of the inter-pret may be compared to the could be of the countries of differences of the acrossic pass, "specifiedly, the state of the countries cunrecross of the network inputs. Specifically, the size of ISD is $3 \times 64 \times 224 \times 224$, whereas that of ResNeX1-101 is $3 \times 16 \times 112 \times 112$. Thus, ISD is 64 times larger than ResNeX1-101. To confirm the accuracies when using larger inputs, we also evaluated the ResNeX1-101 that used $3x64 \times 112 \times 112$ inputs, which are the largest available sizes. in our environment (NVIDIA TITAN X × 8). We can se

Method	Top-1	Top-5	Averag
ResNet-18	54.2	78.1	66.1
ResNet-34	60.1	81.9	71.0
ResNet-50	61.3	83.1	72.2
ResNet-101	62.8	83.9	73.3
ResNet-152	63.0	84.4	73.7
ResNet-200	63.1	84.4	73.7
ResNet-200 (pre-act)	63.0	83.7	73.4
Wide ResNet-50	64.1	85.3	74.7
ResNeXt-101	65.1	85.7	75.4
DenseNet-121	59.7	81.9	70.8
DenseNet-201	61.3	83.3	72.3

Method	Top-1	Top-5	Average
ResNeXt-101	-	-	74.5
ResNeXt-101 (64f)	-	-	78.4
CNN+LSTM [16]	57.0	79.0	68.0
Two-stream CNN [16]	61.0	81.3	71.2
C3D w/ BN [16]	56.1	79.5	67.8
RGB-I3D [3]	68.4	88.0	78.2
Two-stream I3D [3]	71.6	90.0	80.8

4.3. Analyses of fine-tuning

Finally, in this section we confirm the performance of UCF-101 and HMIDB-51. The results of this experiment are important for determining whether the 3D CNNs are effective for other datasets. It should be noted that, in this experiment, fine-mining use only performed to train courts_3 and the fully councered layer became it achieved the best performance during the preliminary experiments.

Table 4 shows the accuracies of Kinetics pertained 3X.

Table 4 shows the accuracies of Kinetics pertained 3X.

Table 4 shows the accuracies of Kinetics pertained 3X.

-18	54.2	78.1	66.1	ResNet-18 (scratch)	42.4
-34	60.1	81.9	71.0		
-50	61.3	83.1	72.2	ResNet-18	84.4
-101	62.8	83.9	73.3	ResNet-34	87.7
152	63.0	84.4	73.7	ResNet-50	89.3
200	63.1	84.4	73.7	ResNet-101	88.9
200	63.0	83.7	73.4	ResNet-152	89.6
-200 (pre-act)				ResNet-200	89.6
esNet-50	64.1	85.3	74.7		_
Ct-101	65.1	85.7	75.4	DenseNet-121	87.6
let-121	59.7	81.9	70.8	ResNeXt-101	50.7
Sec. 201	61.2	93.3	72.2		

trained ResNet-18 clearly outperformed one trained from scratch. This result indicate that pretraining on Kinetics is effective on UCF-101 and HMDB-51. We can also see that effective on UCF-101 and HMDB-51. We can also see that the accuracies bestudyin jumpwood as the depth increased, similar to the results obtained on Kinetics. However, un-like the results on Kinetics, Reselves 200 also improved the accuracies in HMDB-51. Because, as described above, the fine-tuning in this experiment was only performed to tain core-5x and the fully connected layer, the numbers of trained parameters wore the same from Rod-54-50 to ResNet 200. Therefore, the pretrained early layers, which work as feature extractors, relation to the differences of perform of the difference of the difference of the performance of the contraction of the co

datasets. However, the OpensMet-121 results were hower than those of ResNet-0, thereby indicating this tay greater efficiency did not contribute on fine-tening of 3D this greater efficiency did not contribute on fine-tening of 3D this We shows the results of our comparison with state-of-the-art methods in Table 5. Hore, we can see that ResNet/ 101 achieved higher accuracies compared with COD [23], PDD [19], two-sneam CNS [28], and TDD [27]. Further-ison, we can also see that ResNet/Net [101] [27]. Further-tening, we can also see that ResNet/Net [101] [27]. Further-tening, we can also see that ResNet/Net [101] [27]. Further-tening than the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the contribution of the state of the contribution of the contribution of the contribution of the state of the contribution of the contributio

Method	Dim	UCF-101	HME
ResNeXt-101	3D	90.7	63
ResNeXt-101 (64f)	3D	94.5	70
C3D [23]	3D	82.3	
P3D [19]	3D	88.6	
Two-stream I3D [3]	3D	98,0	80
Two-stream CNN [20]	2D	88.0	59
TDD [27]	2D	90.3	63
ST Multiplier Net [7]	2D	94.2	68

that two-stream I3D [3], which utilizes simple two-strear

In this study, we examined the architectures of various CNNs with spatio-temporal 3D convolutional learnets on current video datasets. Based on the results of those experiments, the following conclusions could be obtained: (1) BenNet-18 training resulted in significant overtiting for UCF-101, IMDB-31, and ActivityNet but not for Kinetics. ICC-101, IMSDB-51, and Accomylyche to not for Kinetics, OTI The Kinetics desirate the sufficient after for raining of open 31.72 keys and the resulting of up to 152 kensive discovered to the control of the control of

geNet experienced significant progress in various tasks such as object detection, semantic segmentation, and image cap-tioning. It is felt that, similar to these, 3D CNNs and Kinet-ics have the potential to contribute to significant progress in as have the potential to contribute to significant progress in fields related to various video tasks such as action detection video summarization, and optical flow estimation. In our future work, we will investigate transfer learning not only for action recognition but also for other such tasks.

ECCEPTONCO.
J. S. Arto-El-Baija, N. Kothari, J. Leo, P. Natsor, G. Toderici, B. Warndarajue, and S. Vippomerastenhae. You Tiles-SM: A. Bayes-scale video classification benchmice. arXiv preprint arXiv:1509.098675, 2016. 3
J. Curreira and A. Zhaorettana. Quo vadia, action recognition? A nore model and the Kinetics distored. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (VPPI), pages 4734–4733, 2017. 2.
J. D. Curreira and A. Zhoorettana. Quo vadia, action ecognition?

mAvv: 1903/1700, 2017. 7, 8

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei
ImageNet: A Large-Scale Hierarchical Image Database. In
Proceedings of the IEEE Conference on Computer Vision and

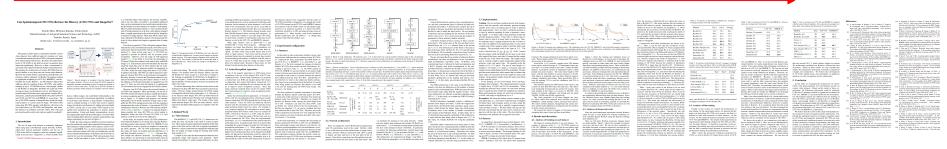
Denoity connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pintrom Recognision (CVPR), pages 4700-4716, 2017, 2, 4, 5, 7 Ill. S. Joffe and C. Strengels. Bachs remailization: Accelerating deep network training by reducing internal convariate shifts. In Proceedings of the International Conference on Manuel Learning, pages 488–496, 2015. 4
S. H., W. Xia, M. Yong, and K. Yu. JD correlational neural

dataset of 101 human action classes from videos in the with CNCV-TR-12-01, 2012. 1, 2, 3, 5 (C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguselos D. Erlan, V. Vanhoudee, and A. Rubtenskin. Going deeper with currectitisms. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), page 1–9, 2015. 3

木を見る(精読)

速読+目的に合わせた読み方を

- 基本的には速読と、部分を詳細に読み込んでいく
- ここから先は何をしたいかに依存する

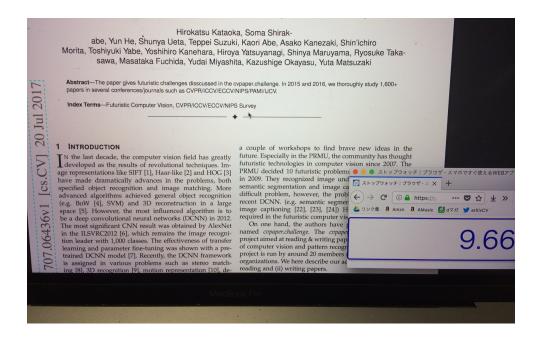


分からない!という場合には最初から最後まで読んで見てもよいかも?

気をつけていること

タイムトライアル

- 時間を気にして、締まりある読みにする
 - 実際ストップウォッチおいて論文読んでます!
- 目安時間
 - 速読(15~30分)
 - 精読(1時間~理解できるまで)



気をつけていること

まとめを作成しよう

- (自分だけでなく他の人が)素早く研究の肝をつかむ
- まとめることで記憶の定着を早くする

[1] SIIN

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, Antonio Torralba, "SUN Database: Large-scale Scene Recognition from Abbey to Zoo", in CVPR2010.

Keywords: Dataset, Scene Categorization, Benchmark, Recognition

概要

<u>データセットの概要</u> シーン認識に関する397クラス 130 51

コンピュータビジョンにおいてシーン認識のデータベースである Scene UNderstanding (SUN) databaseを提案. シーン認識の裾 野を広げた. シーン認識に関する397クラス, 130,519枚の画像が含まれる. 画像例は次ページ. 比較した特徴量は, HOG, denseSIFT, self-similarity (ssim), LBP, GIST, textonなど.

新規性•差分

<u>結果</u>

それまでの物体認識のデータセットでは数百クラスの識別クラスが用意されていたが、シーン認識では15種類程度しか含まれていなかった。SUN databaseでは、それまでのデータセットをさらに拡大させ、397クラスのシーンを含む、大規模なデータセットである。

次ページの図の通り. 全ての特徴量を統合するのが最も精度が高いことが判明した(38.0%). 次いでHOG2x2 (27.2%), geometry texton hist (23.5%), ssim (22.5%), dense SIFT (21.5%)であった.

<u>Links</u>

論文ページ:

http://cs.brown.edu/~havs/papers/sun.pdf

プロジェクトページ:

http://vision.princeton.edu/projects/2010/SUN/

HOG https://hal.archives-ouvertes.fr/inria-00548512/document GIST http://cvcl.mit.edu/scene understanding.html

CIM

http://www.researchgate.net/profile/Eli Shechtman/publication/221362526 Matching Local Self-

Similarities across Images and Videos/links/02e7e520897af25746000000.pdf

DenseSIFT http://www.vision.caltech.edu/Image Datasets/Caltech101/cvpr06b lana.pdf

LBP http://www.outex.oulu.fi/publications/pami 02 opm.pdf

Sparse SIFT http://www.robots.ox.ac.uk/"vgg/publications/papers/sivic04b.pdf">http://www.robots.ox.ac.uk/"vgg/publications/papers/sivic04b.pdf

Texton http://www.ics.uci.edu/~fowlkes/papers/mftm-iccv01.pdf

まとめの例: 概要, 新規性, 手法, 結果, リンク等があるのが望ましい

グループサーベイで意識すること

集団の力をうまく利用する

- 「個人」でのサーベイ
 - ・ 上記の通り
- 「グループ」でのサーベイ
 - 分割と統合をうまく活用して知識を共有
 - 分担して資料を作成してお互いに参照
 - 同じ論文を読んで補完

最近の重要論文は本数が多いので協力して知識を獲得する

ITツールを活用してディスカッション

一人で読むよりも,みんなで協力して進めるのが現代流

- クラウド (Google/Dropbox), チャット (Slack/Skype/Line)
- モチベーション,集合知,作業分割と統合,ディスカッション等グ

ループで活動するメリットは多い



ICCV 2017 速報の資料作成

2. Kd-network

	Erhan, Vincent Vanhoucke, Andrew Rabinovich		Anton van den Hengel, Chris Russell, Anthony Dick, John Bastian, Daniel Pooley, Lachlan Fleming, Lourdes Agapito
<u>1505</u>	Understanding Image Representations by Measuring Their Equivariance and Equivalence Karel Lenc, Andrea Vedaldi	<u>1505</u>	SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite
			Shuran Song, Samuel P. Lichtenberg, Jianxiong Xiao
1505	Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images Anh Nguyen, Jason Yosinski, Jeff Clune	1505	Small-Variance Nonparametric Clustering on the Hypersphere Julian Straub, Trevor Campbell, Jonathan P. How, John W. Fisher III

Monday June 8, 10:10am-12:30pm

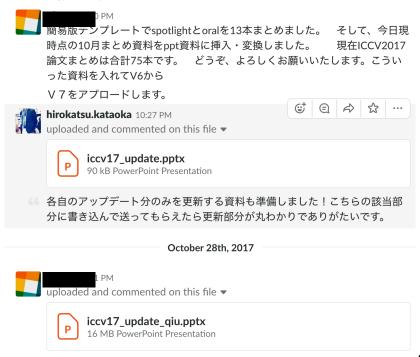
	Poster Session
<u>1505</u>	Going Deeper With Convolutions Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich
<u>1506</u>	Propagated Image Filtering Jen-Hao Rick Chang, Yu-Chiang Frank Wang
<u>1506</u>	Web Scale Photo Hash Clustering on A Single Machine Yunchao Gong, Marcin Pawlowski, Fei Yang, Louis Brandy, Lubomir Bourdey, Rob Fergus
<u>1506</u>	Expanding Object Detector's Horizon: Incremental Learning Framework for Object Detection in Videos Alina Kuznetsova, Sung Ju Hwang, Bodo Rosenhahn, Leonid Sigal
<u>1506</u>	Supervised Discrete Hashing Fumin Shen, Chunhua Shen, Wei Liu, Heng Tao Shen
<u>1505</u>	What do 15,000 Object Categories Tell Us About Classifying and Localizing Actions? Mihir Jain, Jan C. van Gemert, Cees G. M. Snoek
<u>1508</u>	Landmarks-Based Kernelized Subspace Alignment for Unsupervised Domain Adaptation Rahaf Aljundi, Rémi Emonet, Damien Muselet, Marc Sebban

CVPR2015 完全読破

スライドを共有

資料をみんなで作り上げていく

- Ver.を上げていくごとに自分と他人の知見を混ぜていく
- 間接的に読んで、議論を重ねるうちに自分にも知識が定着



気づきはなかなか書きづらいです。ほぼ論文まとめなので、すみません。

CVPR/ICCV 2017 速報の資料作成

可視化,競争

可視化すると意識して読むようになる(?)

- 週間, 月間ランクなど

- 機関総合 (2018/3/1~2018/3/31) 1位:163本,2位:95本,3位:87本

- 個人総合(2018/3/1~2018/3/31) 1位:80本,2位:68本,3位:45本

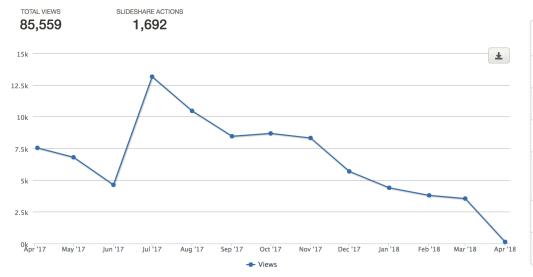
資料公開のススメ

人目に触れて叩き上げる, プレゼンスを高める

- 評価を見て資料の出来栄えを判断(面白いかどうかくらいはわかる)
- 学会などでリアルにあうと反応をもらえる

論文が通った時だけ宣伝, ではなく

普段やっていることでプレゼンスを上げる



Top content			
Name	Views		
CVPR 2017 速報	22,526		
ICCV 2017 速報	9,486		
コンピュータビジョンの今を映す-CVPR 2017 速報より- (夏のトップカンファレンス論文読み会)	4,176		
[2017.03] cvpaper.challenge2017	2,649		
CVPR 2016 まとめ v1	2,315		

サーベイ法まとめ

サーベイの方法論について,個人/グループという単位で説明

- 速読と精読を組み合わせた個人サーベイ
- 組織的サーベイで早く/確実に知識を作り上げていく

サーベイ(に限らず研究)は楽 しんでやるもの!こんな楽しい ことができるようになった,分 からなかったことを知識として 明らかにした, というゲーム

MIRU若手Pにご参加ください!

https://sites.google.com/view/miru2018sapporo/wakate_top

今回のテーマは「異分野サーベイ」

■ 3/18(日): 募集開始

■ 4/23(月): 募集締切

■ 4/27(金): グループ分けの発表,企画1: 異分野サーベイ企画開始.

■ 5/17(木)-18(金): サーベイに関するチュートリアル講演(PRMU研究会@岐阜大学).

■ スケジュールの詳細は<u>PRMU</u> のウェブサイトに掲載予定です。また、本講演の様子や資料は<u>企</u> <u>画内容のページ</u>にアップロード予定です。

