論文紹介

Action Recognition in Videosについて調査

- 動画認識
  - 動画から人が何をしているか分類する

Two-Stream Convolutional Networks for Action Recognition in Videos

(2014) Simonyan et. al,
https://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.pdf

どんなもの？

Two stream ConvNet architecutreを提案

どうやって有効だと検証した？

Table 4: **Mean accuracy (over three splits) on UCF-101 and HMDB-51.**

Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

技術の手法や肝は？

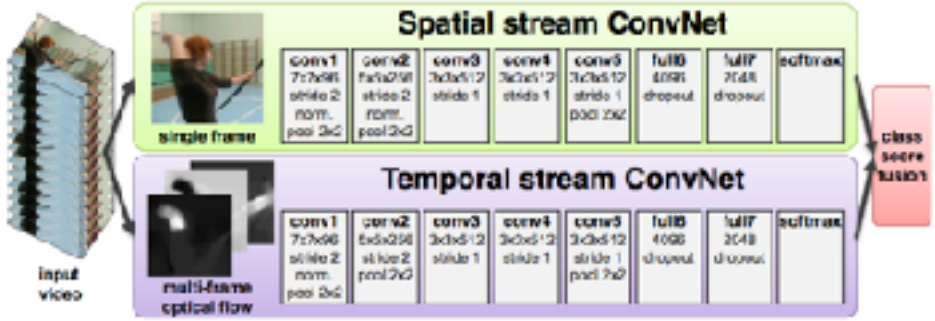


Figure 1: **Two-stream architecture for video classification.**

議論はある？

膨大な量のトレーニングデータ(multiple TABS)について、大きな課題となる。

先行研究と比べて何がすごい？

Spatial(空間的)な情報とTemporal(時間的)な情報を統合した点

*これまではTemporalのみ

次に読むべき論文は？

***

Learning Spatiotemporal Features with 3D Convolutional Networks

(2015) Tran et. al,
https://arxiv.org/abs/1412.0767

どんなもの？

C3Dを提案

どうやって有効だと検証した？

Dataset Task	Sport1M action recognition	UCF101 action recognition	ASLAN action similarity labeling	YUPENN scene classification	UMD scene classification	Object object recognition
Method	[29]	[39]([25])	[31]	[9]	[9]	[32]
Result	**90.8**	75.8 (89.1)	68.7	96.2	77.7	42.0
C3D	85.2	**85.2 (90.4)**	**78.3**	**98.1**	**87.7**	**22.3**

Table 1. **C3D compared to best published results.** C3D outperforms all previous best reported methods on a range of benchmarks except for Sports-1M and UCF101. On UCF101, we report accuracy for two groups of methods. The first set of methods use only RGB frame inputs while the second set of methods (in parentheses) use all possible features (e.g. optical flow, improved Dense Trajectory).

技術の手法や肝は？

Figure 3. **C3D architecture.** C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from pool1 to pool5. All pooling kernels are $2 \times 2 \times 2$, except for pool1 is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

議論はある？

3Dなので、パラメータを最適化するために十分なデータセットが必要。

→UCF101についてTwo-Stream CNNより劣る

先行研究と比べて何がすごい？

3Dの畳み込みを実施したので、時間的な情報を失わない点。

次に読むべき論文は？

***

Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks

(2017) Simonyan Qiu et. al,
https://arxiv.org/pdf/1711.10305.pdf

どんなもの？

P3Dを提案

どうやって有効だと検証した？

Method	Accuracy
End-to-end CNN architecture with fine-tuning	
Two-stream ConvNet [25]	73.0% (88.0%)
Factorized ST-ConvNet [29]	71.3% (88.1%)
Two-stream + LSTM [37]	82.6% (88.6%)
Two-stream fusion [6]	82.6% (92.5%)
Long-term temporal ConvNet [33]	82.4% (91.7%)
Key-volume mining CNN [39]	84.5% (93.1%)
ST-ResNet [4]	82.2% (93.4%)
TSN [36]	85.7% (94.0%)
CNN-based representation extractor + linear SVM	
C3D [31]	82.3%
ResNet-152	83.5%
P3D ResNet	**88.6%**
Method fusion with IDT	
IDT [34]	85.9%
C3D + IDT [31]	90.4%
TDD + IDT [35]	91.5%
ResNet-152 + IDT	92.0%
P3D ResNet + IDT	**93.7%**

技術の手法や肝は？

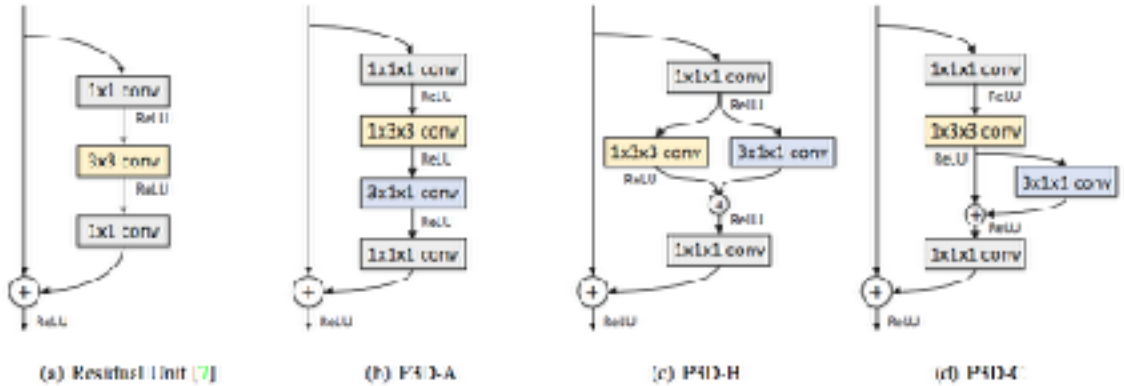


Figure 3. Bottleneck building blocks of Residual Unit and our Pseudo-3D.

議論はある？

大規模なデータセットでの学習が必要

先行研究と比べて何がすごい？

3x1x1と1x3x3の畳み込みを実施することで、擬似的な3次元の畳み込みを実施した点。また、パラメータ数を削減した点

次に読むべき論文は？

***

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

(2017) Carreira et. al
https://arxiv.org/pdf/1705.07750.pdf

どんなもの？

データセット(Kinetics)の提案

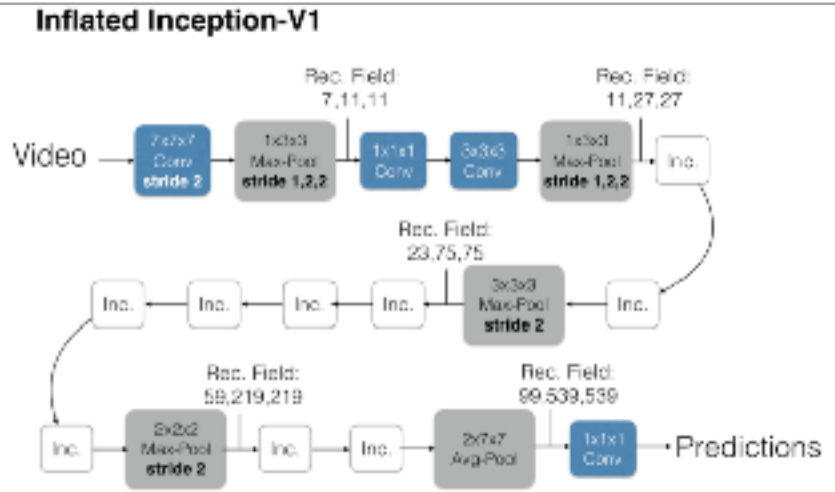
→state-of-the-artのモデルを再評価する。

また、I3Dを提案する。

どうやって有効だと検証した？

Model	UCF-101	HMDB-51
Two-Stream [27]	88.0	59.4
IDT [33]	86.4	61.7
Dynamic Image Networks + IDT [2]	89.1	65.2
TDD + IDT [34]	91.5	65.9
Two-Stream Fusion + IDT [8]	93.5	69.2
Temporal Segment Networks [35]	94.2	69.4
ST-ResNet + IDT [7]	94.6	70.3
Deep Networks [15], Sports 1M pre-training	65.2	-
C3D one network [31], Sports 1M pre-training	82.3	-
C3D ensemble [31], Sports 1M pre-training	85.2	-
C3D ensemble + IDT [31], Sports 1M pre-training	90.1	-
RGB-I3D, Imagenet+Kinetics pre-training	95.6	74.8
Flow-I3D, Imagenet+Kinetics pre-training	96.7	77.1
Two-Stream I3D, Imagenet+Kinetics pre-training	**98.0**	80.7
RGB-I3D, Kinetics pre-training	95.1	74.3
Flow-I3D, Kinetics pre-training	96.5	77.3
Two-Stream I3D, Kinetics pre-training	97.8	**80.9**

技術の手法や肝は？



議論はある？

このモデルは、先行研究の包括的なモデル、人間の行動に注目したaction tuesやattention mechanismsなどを適用してない。

先行研究と比べて何がすごい？

これまでの3DCNNは、学習データが少ないため精度が出ていなかったが、新しいデータ(Kinetics)により制度が出た

次に読むべき論文は？

attention mechanismsとは?

VideoLSTM Convolves, Attends and Flows for Action Recognition

(2016) Zhenyang Li et. al
https://arxiv.org/pdf/1607.01794.pdf

どんなもの？

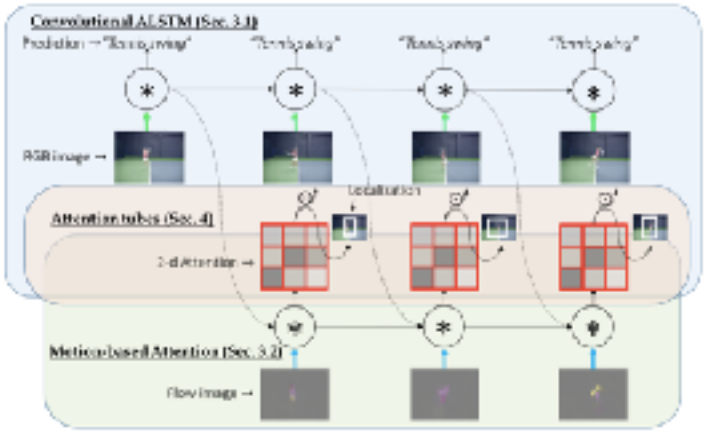
VideoLSTM (a new recurrent neural network architecture)の提案

どうやって有効だと検証した？

		ConvNet	LSTM	ALSTM	ConvLSTM	ConvALSTM
UCF101	RGB Appearance	77.4	77.5	77.0	77.6	79.6
	Optical flow	75.2	78.3	79.5	80.4	82.1
HMDB51	RGB Appearance	42.2	41.3	40.9	41.8	43.3
	Optical flow	41.8	46.0	49.2	48.2	52.6

Table 1: **Convolutional ALSTM networks**. For both appearance and optical flow input, our proposed ConvALSTM improves accuracy the most.

技術の手法や肝は？



議論はある？

なし

先行研究と比べて何がすごい？

Attentionの情報とRGBの情報を組み合わせた。

→シンプルなConvNetにattetntionの情報を組み合わせることにより、精度が向上した

次に読むべき論文は？

***