UNIVERSITÀ DI PISA

A.Y. 2019/20
« Intelligent systems for pattern recognition »
Master Degree in Computer Science
Artificial Intelligence Curriculum

Convolutional NN for video processing

Midterm 4

Diletta Goglia

Problem

- Learning spatio-temporal video representation: video is a temporal sequence of frames with large variations and complexities, resulting in difficulty in learning.
- Performing 3D convolutions to capture both spatial and temporal dimensions in videos is expensive in computational cost and memory demand.
- Training of 3D CNN is very computationally expensive, and the model size also has a quadratic growth compared to 2D CNN making it extremely difficult to train a very deep 3D CNN.

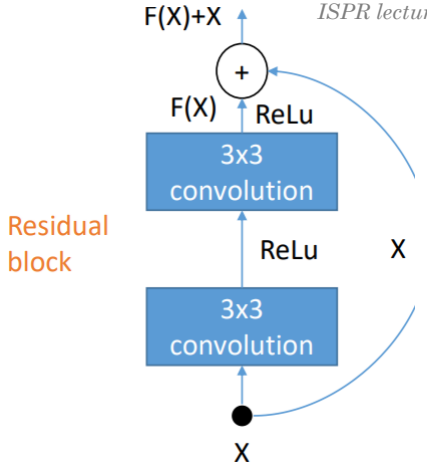
Solution

- Devising a family of bottleneck building blocks that leverages both spatial and temporal convolutional filters
- Simulating 3x3x3 convolutions with 1x3x3 convolutional filters on spatial domain (equivalent to 2D CNN) plus 3x1x1 convolutions to construct temporal connections on adjacent feature maps in time
- ResNet-like architecture

References

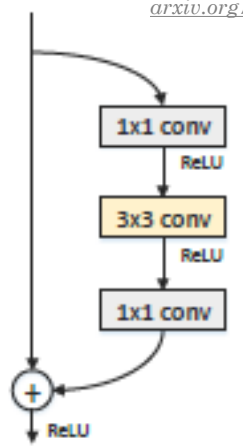
ResNet (2015) – Residual Blocks

ISPR lecture 13 (17/04/20)



Bottleneck architecture

arxiv.org/abs/1711.10305 , pag.4



3D convolutions

3D convolutional filters has size *d x k x k* , where d is the temporal depth of kernel and k is the kernel spatial size

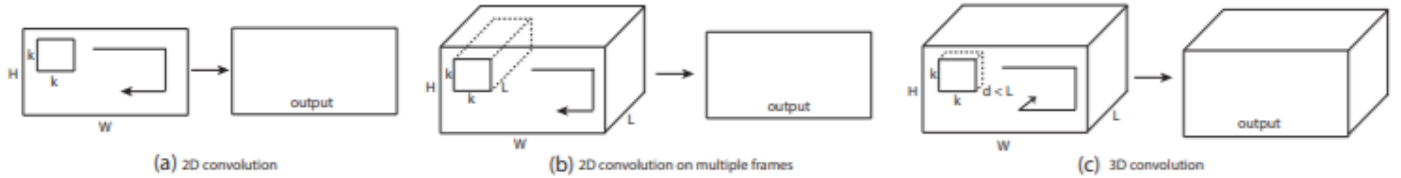


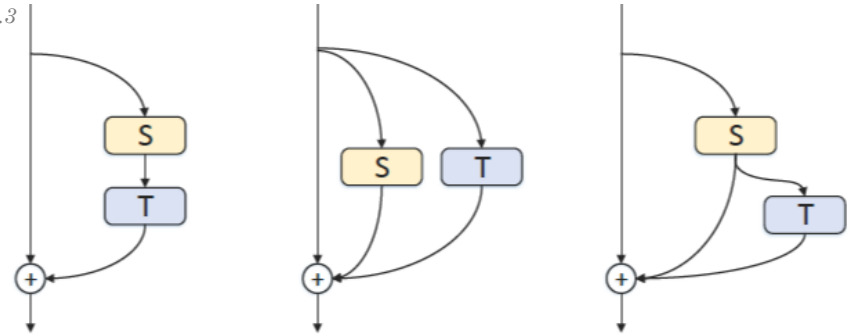
Figure 1. 2D and 3D convolution operations. a) Applying 2D convolution on an image results in an image. b) Applying 2D convolut on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume resu in another volume, preserving temporal information of the input signal.

Du Tran et al., *Learning Spatiotemporal Features with 3D Convolutional Networks* 2015, https://arxiv.org/pdf/1412.0767.pdf

Model description

1) P3D blocks

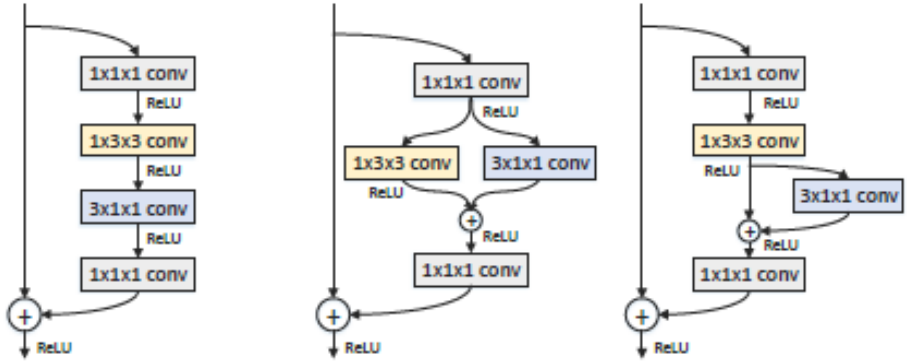
arxiv.org/abs/1711.10305 , pag.3



Blocks design	P3D-A	P3D-B	P3D-C
Influence of spatial (S) and temporal (T) dimensions	Direct influence (cascaded manner)	Indirect influence (parallel fashion)	a compromise between P3D-A and P3D-B
Other considerations	Only the temporal 1D filters are directly connected to the final output	Both filters are at different pathways and both are directly accumulated into the final output	simultaneously building the direct influences among S, T and the final output

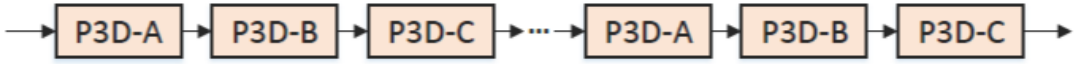
2) Bottleneck architecture

arxiv.org/abs/1711.10305 , pag.4



3) Complete model

arxiv.org/abs/1711.10305 , pag.5



Mixing different P3D Blocks. "*Further inspired from the recent success of pursuing structural diversity in the design of very deep networks*" → complete version of P3D ResNet by mixing different P3D blocks in the architecture to enhance structural diversity replace Residual Units with a chain of P3D blocks in the order P3D-A, P3D-B, P3D-C.

Considerations

1) Replace 2D Residual Units in ResNet with a new family of building blocks: Pseudo-3D (P3D), to leverage both spatial and temporal convolutional filters.
2D kernel to encode spatial information + 1D kernel for temporal dimension
Reduces the model size significantly, but also enables the pre-training of 2D CNN from image data, endowing Pseudo 3D CNN more power of leveraging the knowledge of scenes and objects learnt from images.

2) Modified with a **bottleneck design** for reducing the computation complexity: the first and last 1x1 convolutional layers are applied for reducing and restoring dimensions of input sample, respectively.

3) A novel **Pseudo-3D Residual Net** (P3D ResNet) is developed, composing each P3D block at different placement in ResNet-like architecture

Comparison 1

- First comparison is conducted on UCF101 video action recognition dataset (see: "K. Soomro et al. , *UCF101: A dataset of 101 human action classes from videos in the wild.*")
- All the three P3D ResNet variants exhibit better performance than ResNet-50
- The results basically indicate the advantage of exploring spatio-temporal information by P3D blocks.

Table 1. Comparisons of ResNet-50 and different Pseudo-3D ResNet variants in terms of model size, speed, and accuracy on UCF101 (split1). The speed is reported on one NVidia K40 GPU.

Method	Model size	Speed	Accuracy
ResNet-50	92MB	15.0 frame/s	80.8%
P3D-A ResNet	98MB	9.0 clip/s	83.7%
P3D-B ResNet	98MB	8.8 clip/s	82.8%
P3D-C ResNet	98MB	8.6 clip/s	83.0%
P3D ResNet	98MB	8.8 clip/s	84.2%

Comparison 2

- Learning of P3D ResNet: on Sports-1M dataset (about 1.13 million videos annotated with 487 Sports labels)
- Followed the official split: 70% training, 10% validation, 20% test.
- Measuring video/clip classification accuracy on the test set. Randomly sample 20 clips from each video and adopt a single center crop per clip, which is propagated through the network to obtain a clip-level prediction score. The video-level score is computed by averaging all the clip-level scores of a video.

Comparing the following approaches for performance evaluation:

Table 2. Comparisons in terms of pre-train data, clip length, Top-1 clip-level accuracy and Top-1&5 video-level accuracy on Sports-1M.

Method	Pre-train Data	Clip Length	Clip hit@1	Video hit@1	Video hit@5
Deep Video (Single Frame) [10]	ImageNet1K	1	41.1%	59.3%	77.7%
Deep Video (Slow Fusion) [10]	ImageNet1K	10	41.9%	60.9%	80.2%
Convolutional Pooling [37]	ImageNet1K	120	70.8%	72.3%	90.8%
C3D [31]	–	16	44.9%	60.0%	84.4%
C3D [31]	I380K	16	46.1%	61.1%	85.2%
ResNet-152 [7]	ImageNet1K	1	46.5%	64.6%	86.4%
P3D ResNet (ours)	ImageNet1K	16	47.9%	66.4%	87.4%

P3D ResNet leads to a performance boost against ResNet-152 (2D CNN) and C3D (3D CNN) by 1.8% and 5.3% in terms of top-1 video-level accuracy, respectively. The results basically indicate the advantage of exploring spatio-temporal information by decomposing 3D learning into 2D convolutions in spatial space and 1D operations in temporal dimension.
P3D ResNet is benefited from the principle of structural diversity in network design.

Since "*the networks could be utilized as a generic representation extractor for any video analysis tasks*"
→ Next slides: evaluation of P3D ResNet video representation on three different tasks and five popular datasets.

Evaluation 1

1. Comparison with several state-of-the-art techniques in the context of video action recognition (UCF 101 dataset)

- Most recent CNN architectures often employ and fuse two or multiple types of inputs, e.g., frame or audio. Hence, the performances by exploiting only video frames and late fusing two kind of inputs are both reported (the second in brackets). This second method leads to apparent improvement compared to only using video frames. "*This motivates us to learn P3D ResNet architecture with other types of inputs in our future works.* "

- Furthermore, P3D ResNet utilizing 2D spatial convolutions plus 1D temporal convolutions exhibits significantly better performance than C3D which directly uses 3D spatio-temporal convolutions.

- By combining with IDT which are hand-crafted features, the performance will boost up to 93.7%.

- Compared to which operates LSTM over high-level representations of frames to explore temporal information, P3D ResNet is benefited from the temporal connections throughout the whole architecture and outperforms.

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%**

Evaluation 2

2. Action similarity labeling challenge on ASLAN benchmark: answer a binary question of "does a pair of videos present the same action?"

- The video-level representation is obtained by averaging all clip-level representations. Given each video pair, calculate 12 different similarities on each type of video representation and then a binary classifier is trained.

- Unlike the observations on action recognition task, C3D significantly outperforms ResNet-152 on the scenario of action similarity labelling: this may be the result of difficulty in interpreting the similarity between videos based on the ResNet-152 model learnt purely on image domain. In contrast, the video representation extracted by C3D which is trained on video data potentially has higher capability to distinguish between videos.

- Improvements are also observed in P3D ResNet → advantages of both C3D and ResNet-152 by pre-training 2D spatial convolutions on image data and learning 1D temporal connections on video data.

Method	Model	Accuracy	AUC
STIP [13]	linear	60.9%	65.3%
MIP [12]	metric	65.5%	71.9%
IDT+FV [19]	metric	68.7%	75.4%
C3D [31]	linear	78.3%	86.5%
ResNet-152 [7]	linear	70.4%	77.4%
P3D ResNet	linear	**80.8%**	**87.9%**

Evaluation 3

3. The third experiment was conducted on scene recognition. Table 6 shows the accuracy of different methods.

Table 6. The accuracy performance of scene recognition on Dynamic Scene and YUPENN sets.

Method	Dynamic Scene	YUPENN
[3]	43.1%	80.7%
[5]	77.7%	96.2%
C3D [31]	87.7%	98.1%
ResNet-152 [7]	93.6%	99.2%
P3D ResNet	**94.6%**	**99.5%**

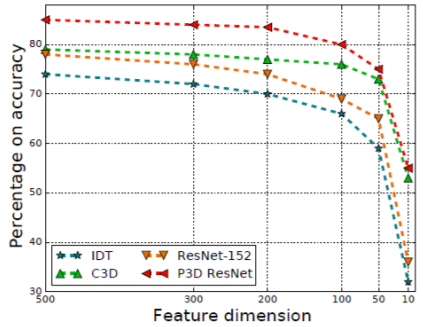


Figure 6. The accuracy of video representation learnt by different architectures with different dimensions. The performances reported in this figure are on UCF101 (3 splits).

Figure 6 compares the accuracy of video representation with different dimensions on UCF101 dataset.
Overall, video representation learnt by P3D ResNet consistently outperforms others at each dimension. In general, higher dimensional representation provide better accuracy.

P3D ResNet is benefited from the exploration of knowledge from both image and video domain, making the learnt video representation more robust to the change of dimension.

Visualization

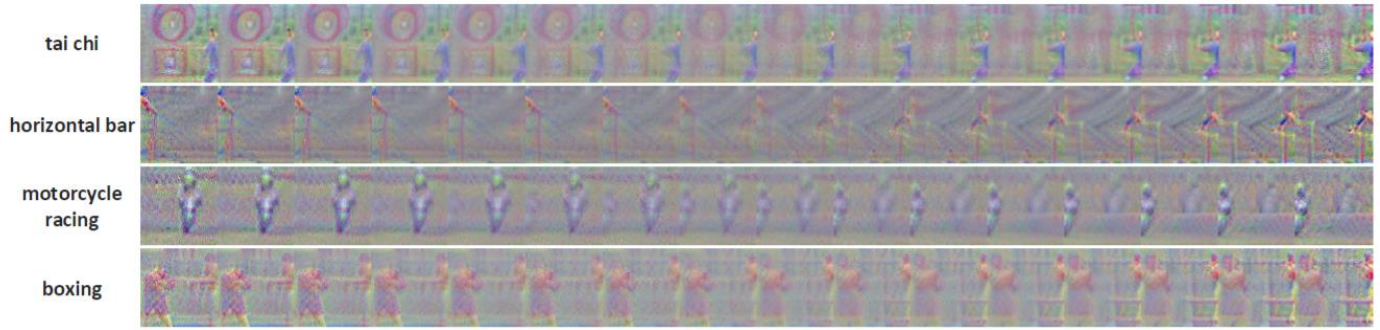
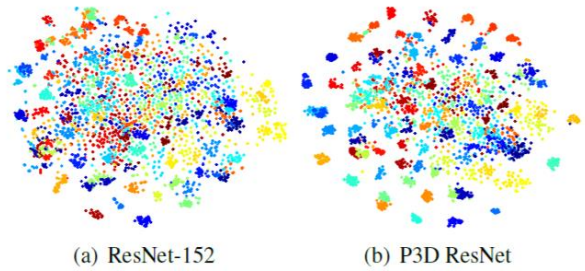


Figure 5. Visualization of class knowledge inside P3D ResNet model by using DeepDraw [1]. Four categories, i.e., tai chi, horizontal bar, motorcycle racing and boxing, are selected for visualization.

Visualization of class knowledge inside P3D ResNet (with DeepDraw toolbox).
P3D ResNet model could capture both spatial visual patterns and temporal motion. In the tai chi example, the model generates a video clip in which a person is displaying different poses, depicting the process of this action.



(a) ResNet-152 (b) P3D ResNet

Figure 7. Video representation embedding visualizations of ResNet-152 and P3D ResNet on UCF101 using t-SNE [32]. Each video is visualized as one point and colors denote different actions.

T-SNE visualization of embedding of video representation learnt by ResNet-152 and P3D ResNet.
Randomly select 10K videos from UCF101 and the video-level representation is then projected into 2-dimensional space using t-SNE. It is clear that video representations by P3D ResNet are better semantically separated than those of ResNet-152.

Conclusions and personal considerations

Overall view

- Experiments conducted on five datasets in the context of video action recognition, action similarity labeling and scene recognition demonstrate the effectiveness and generalization of the spatio-temporal video representation produced by P3D ResNet. Performance improvements are clearly observed when comparing to other feature learning techniques.

Strong points

- Reduced computational complexity (less dimensional kernels: from 3D to 2D+1D)
- Bottleneck architecture to enforce dimensionality reduction

Novelties

- "*By additionally pursuing structural diversity, P3D ResNet makes the absolute improvement over P3D-A ResNet, P3D-B ResNet and P3D-C ResNet by 0.5%, 1.4% and 1.2% in accuracy respectively, indicating that enhancing structural diversity with going deep could improve the power of neural networks*"

→ more a confirmation than a result : "X. Zhang et al., *Polynet: A pursuit of structural diversity in very deep networks*, 2016".

Future improvements

- Extend P3D ResNet learning to other types of inputs, like audio.

UNIVERSITÀ DI PISA

A.Y. 2019/20
« Intelligent systems for pattern recognition »
Master Degree in Computer Science
Artificial Intelligence Curriculum

Thanks for your attention