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

Dawit Anelay
d.anelay1@studenti.unipi.it

Intelligent Systems for Pattern Recognition

Introduction to the problem

300 hours of video are uploaded to YouTube every minute!

- **CNN** • For Image Processing problem . It's has been a powerful approach model.
  - • For Learning Spatio-Temporal Representation ? For Video Processing,? E.g For Human Action detection ?.
- **A Rewarded Study Approche** • Performing 3D Convolutions (Encourage Performance on Sport-1M dataset 85.2%).
  - • Model : C3D,ResNet-152 .
- **Problem** • **Computational Cost** and **Memory Demand** . (Model C3D--321MB with 11-layer, ResNet-152--235MB with 152-layer).
  - • Model Size has **Quadratic Growth** Compare to 2D CNN AND **Very Difficult to Train** Because they are very deep.
- **Proposed Idea** • Why not recycle off-the-shelf 2D networks for a 3D CNN.
- **Solution** • **De-Couple 3D CNN** =**2D CNN** (**Spatio-Domain**) + **1D CNN** (**Temporal -Domain**).
  - • A New Architecture Named **P3DResNet** with a new designed Block model that simulate 3D CNN in an **Economic** and **efficient** way.

P3D Blocks and P3D ResNet

- Design issue while decoupling 3D CNN.

① The first issue is about whether the modules of 2D filters on spatial dimension (S) and 1D filters on temporal domain (T) should directly or indirectly influence each other.

‣Directly - cascaded manner ‣ Indirectly-parallel fashion.

② The second issue is whether the two kinds of filters should both directly influence the final output.

(1) P3D-A:The two kinds of filters can directly influence each other in the same path and only the temporal 1D filters are directly connected to the final output,

- **$(I + T \cdot S) \cdot x_t := x_t + T(S(x_t)) = x_{t+1}$.**

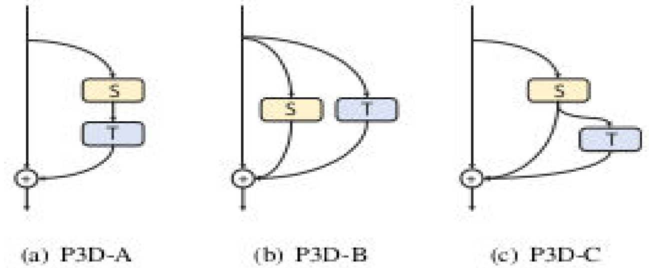

Figure 2. Three designs of Pseudo-3D blocks.

(2) P3D-B: There is no direct influence between S and T, both of them are directly accumulated into the final output

- **$(I + S + T) \cdot x_t := x_t + S(x_t) + T(x_t) = x_{t+1}$.**

(3) P3D-C:A compromise between P3D-A and P3D-B, by simultaneously building the direct influences among S, T and the final output.

- **$(I + S + T) \cdot x_t := x_t + S(x_t) + T(x_t) = x_{t+1}$.**

Bottleneck architectures.

• Three P3D ResNet variants, i.e., P3D-AResNet, P3D-B ResNet and P3D-C ResNet by replacing all the Residual Units in a 50-layer ResNet (ResNet-50) with one certain kind of P3D block, respectively.

• A complete version of P3D ResNet is proposed by mixing all the three P3D blocks from the viewpoint of structural diversity.Residual Units with a chain of

Figure 4. P3D ResNet by interleaving P3D-A, P3D-B and P3D-C.

P3D blocks in the order P3D-A→P3D-B→P3D-C.

• No explicit reason stated why this type of order.

• The speed of the model is very fast and could reach 9 clips per second.

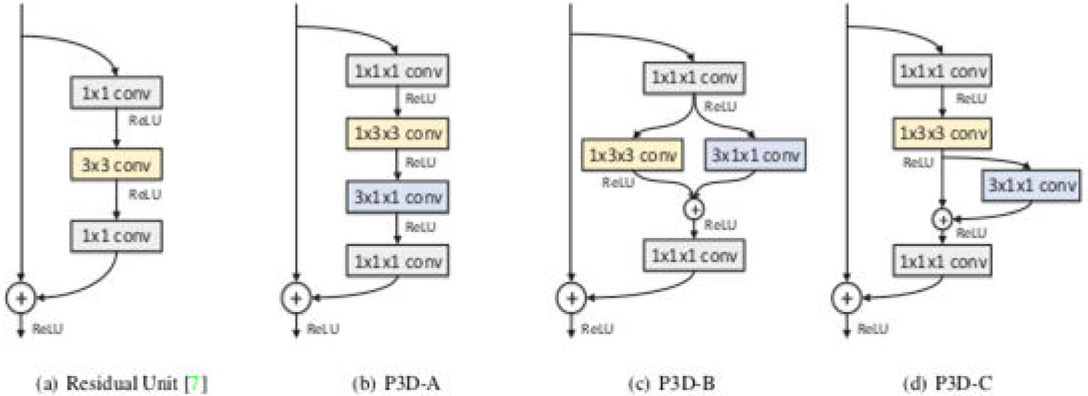

Figure 3. Bottleneck building blocks of Residual Unit and our Pseudo-3D.

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%

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

Spatio-Temporal Representation Learning-Result

- The learning conducted on Sports-1M dataset(1.13 million videos annotated with 487 Sports labels).
- 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.

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]	1380K	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%

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.

Video Representation Evaluation on three different tasks and five popular datasets

① action recognition - UCF101 and ActivityNet dataset

②action similarity - does a pair of videos present the same action?"-ASLAN

③ scene recognition on Dynamic Scene and YUPENN sets.

- In all task the model Leads to a performance boost

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%**

Table 5. Action similarity labeling performances on ASLAN benchmark. STIP.STIP: Space-Time Interest Points; MIP: Motion Interchange Patterns; FV: Fisher Vector.

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%**

Conclusion

- **Strong points**

① P3D ResNet is an effect way for For Learning Spatio-Temporal Representation .

② Performance improvements are clearly observed when comparing to other feature learning techniques.

- **weaknesses**

① The speed of the model is consider only with a small video frame size (16 frame -long).

- **Futurework**

① Attention mechanism

② Extend P3D ResNet learning to other types of inputs, e.g., optical flow or audio.