Exploring Temporal Differences in 3D Convolutional Neural Networks

Gagan Kanojia, Sudhakar Kumawat, Shanmuganathan Raman

Indian Institute of Technology Gandhinagar, India

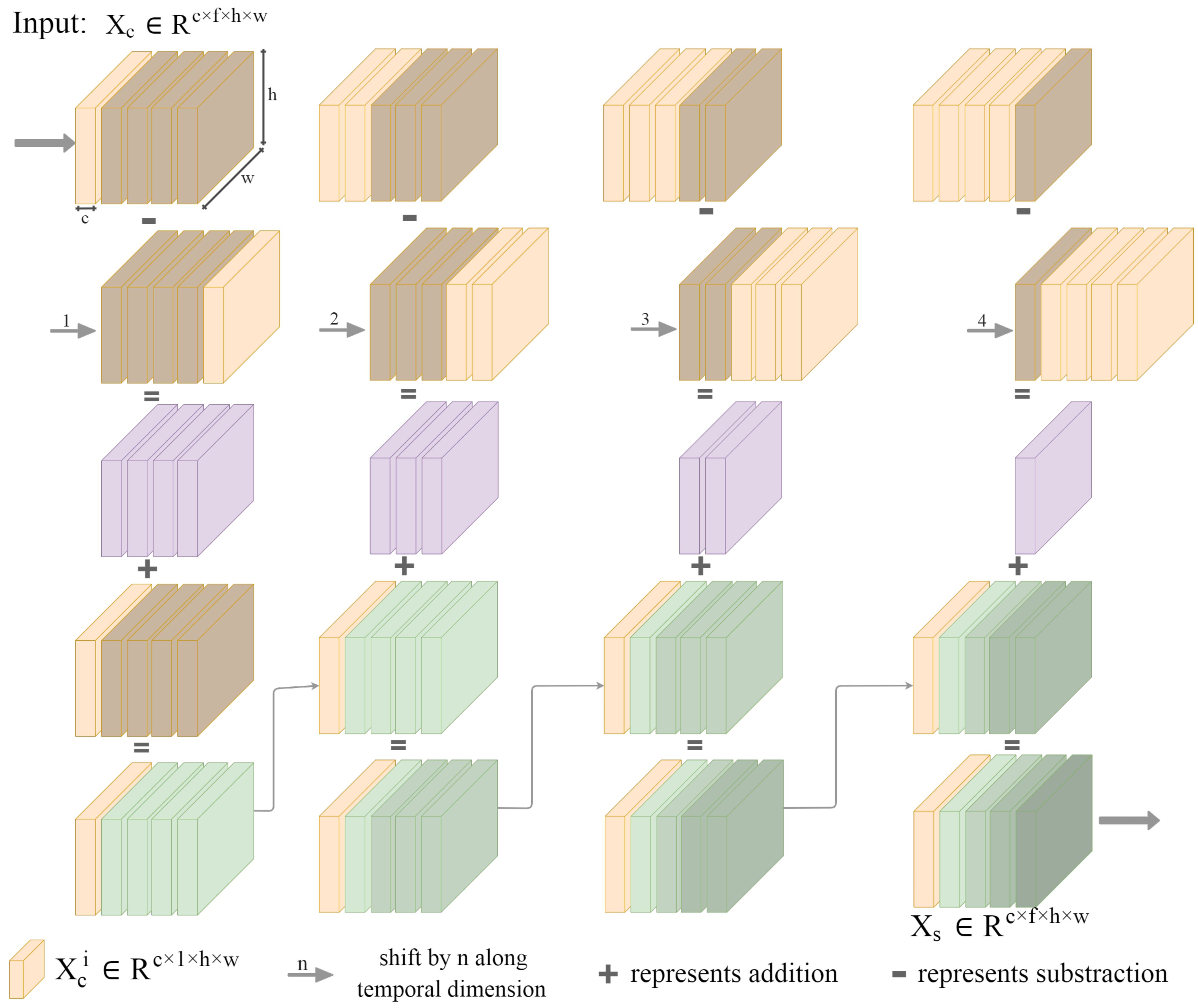
Problem Definition and Contribution

Goal: Capturing both spatial and temporal features of the 3D data while reducing the cost in terms of trainable parameters.

Key Contributions: We propose a novel convolutional block which can serve as an alternative to standard 3D convolutional layer.

- The proposed convolutional block captures spatial information by performing 2D convolution and temporal information using simple operations of shift, subtract and add.
- We reduce the number of parameters by a factor of n by replacing the 3D convolution kernel of size $n \times n \times n$ with the proposed convolution block.
- We show that the proposed convolutional block helps the 3D CNNs to perform better while utilizing lesser parameters than the standard 3D convolution kernels.

SSA Layer


Input: $X_c \in R^{c \times f \times h \times w}$
h
w
c
1
2
3
4
$X_s \in R^{c \times f \times h \times w}$
$X_c^i \in R^{c \times 1 \times h \times w}$
n
shift by n along temporal dimension
+ represents addition
- represents substraction

Experiments & Results on UCF101

Network	Layers	Parameters (Millions)	SSA Layer	Temporal pooling	Accuracy(%)
3D ResNet (baseline)	18	≈ 33			45.6
SSA-ResNet (ours)	18	≈ 11		✓	52.8
SSA-ResNet (ours)	18	≈ 11	✓	✓	**55.7**
3D ResNeXT (baseline)	50	≈ 26			49.3
SSA-ResNeXT (ours)	50	≈ 23		✓	54.9
SSA-ResNeXT (ours)	50	≈ 23	✓	✓	**56.9**
3D WideResNet (baseline)	50	≈ 157			46.8
SSA-WideResNet(ours)	50	≈ 67		✓	50.7
SSA-WideResNet(ours)	50	≈ 67	✓	✓	**52.9**
C3D (baseline)	5	≈ 18			44
SSA-C3D (ours)	5	≈ 14		✓	50
SSA-C3D (ours)	5	≈ 14	✓	✓	**51.6**
3D ResNet (baseline)	101	≈ 88			46.7
SSA-ResNet (ours)	101	≈ 43		✓	52.1
SSA-ResNet (ours)	101	≈ 43	✓	✓	**54.4**

Comparisons with baselines. The comparison of the test accuracies obtained by the baseline 3D models with the networks obtained by replacing the 3D convolution kernel by the proposed convolution block in the baseline 3D models on UCF101 split-1 when trained from scratch.

Network	Layers	Parameters (Millions)	Model Size (MB)	Accuracy
2D-ResNet	18	≈11.2	-	42.2
2D-ResNet	34	≈21.5	-	42.2
3D-ResNet	18	≈33.2	254	45.6
3D-ResNet	34	≈63.5	485	45.9
3D-ResNet	101	≈86.06	657	46.7
3D STC-ResNet	18	-	-	42.8
3D STC-ResNet	50	-	-	46.2
3D STC-ResNet	101	-	-	47.9
C3D	5	≈18	139.6	44
R(2+1)D	18	≈33.3	128	48.37
SSA-ResNet (ours)	18	≈11	88.5	55.7
SSA-ResNeXt (ours)	50	≈23	185.9	**56.9**

Comparisons with the state-of-the-art. The comparison of the proposed approach with the state-of-the-art methods when trained from scratch on UCF101 dataset.

#Shift	Temporal pooling	Accuracy
0		46.3
0	✓	52.8
1	✓	52.6
2	✓	53.4
3	✓	53.9
f-1		51.3
f-1	✓	**55.7**

Analysis of different shifts and temporal pooling. The test accuracies obtained using SSA-ResNet (18 layers) with varying number of shifts along with the effect of temporal pooling.

Formulation

The proposed convolutional block has three parts: 2D convolution kernel, SSA layer, and temporal pooling layer.

2D convolution. In the proposed framework, first we obtain $\mathcal{X}_c = \mathcal{X} \star g$. Here, $\star$ stands for convolution, and g is a 2D filter of kernel size $1 \times k \times k$ and c channels.

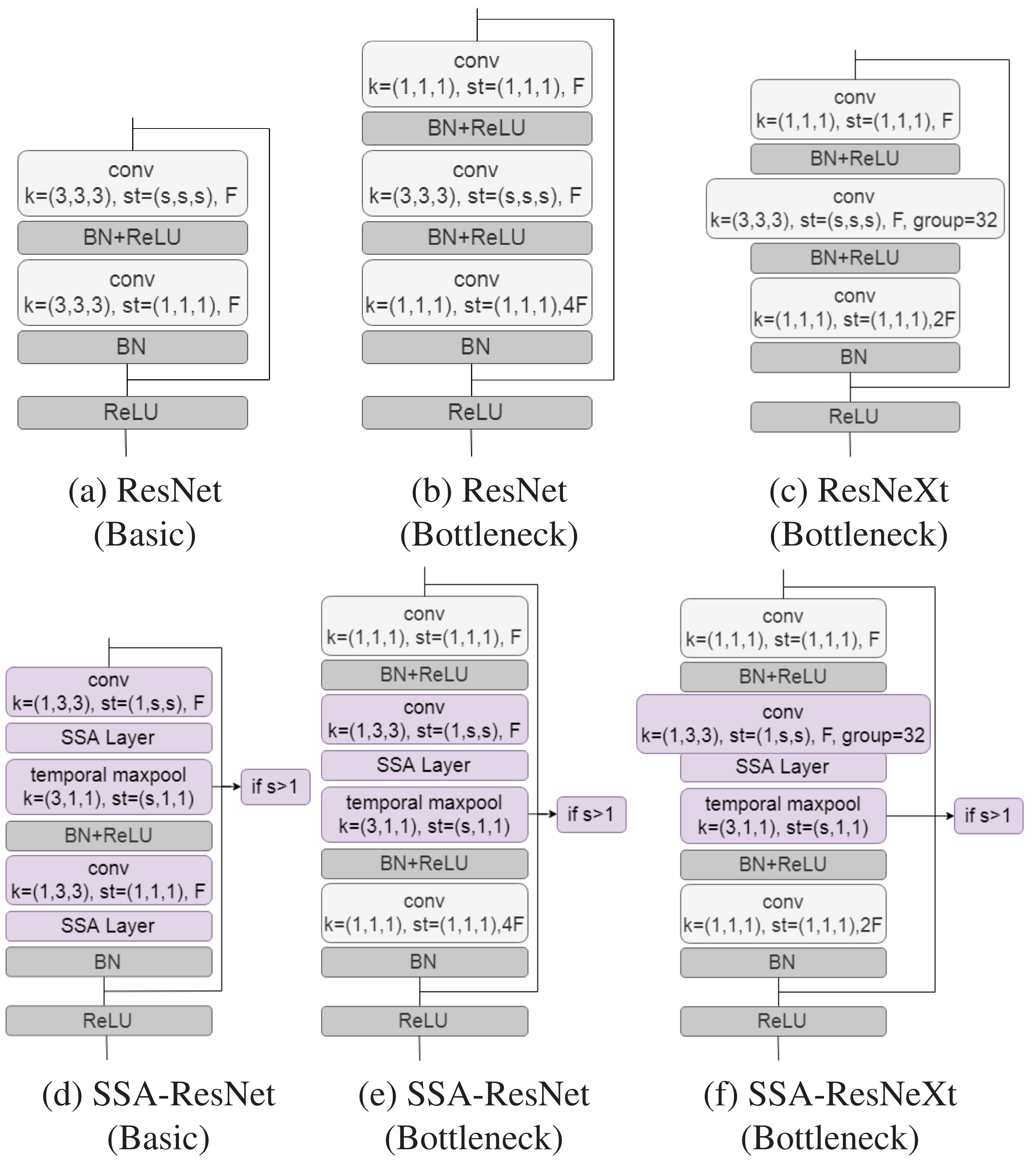
SSA Layer. SSA stands for Shift, Subtract and Add operations performed in SSA layer. The purpose of SSA layer is to extract the temporal information present in the spatio-temporal data.

$$\mathcal{X}_s^i = \mathcal{X}_c^i + \frac{1}{f}\sum_{k=1}^{i-1}\frac{f-(i-k)}{f}(\mathcal{X}_c^i - \mathcal{X}_c^k), \quad \forall i = 2, \dots, f \quad (1)$$

Here, k is the shift and i is temporal location.

Temporal pooling. In our case, we are not performing convolution along the temporal depth. Hence, to reduce the temporal depth, we perform max pooling along the temporal direction of the feature maps.

Network Architecture


conv k=(3,3,3), st=(s,s,s), F
BN+ReLU
conv k=(3,3,3), st=(1,1,1), F
BN
ReLU

(a) ResNet (Basic)

conv k=(1,1,1), st=(1,1,1), F
BN+ReLU
conv k=(3,3,3), st=(s,s,s), F
BN+ReLU
conv k=(1,1,1), st=(1,1,1),4F
BN
ReLU

(b) ResNet (Bottleneck)

conv k=(1,1,1), st=(1,1,1), F
BN+ReLU
conv k=(3,3,3), st=(s,s,s), F, group=32
BN+ReLU
conv k=(1,1,1), st=(1,1,1),2F
BN
ReLU

(c) ResNeXt (Bottleneck)

conv k=(1,3,3), st=(1,s,s), F
SSA Layer
temporal maxpool k=(3,1,1), st=(s,1,1) → if s>1
BN+ReLU
conv k=(1,3,3), st=(1,1,1), F
SSA Layer
BN
ReLU

(d) SSA-ResNet (Basic)

conv k=(1,1,1), st=(1,1,1), F
BN+ReLU
conv k=(1,3,3), st=(1,s,s), F
SSA Layer
temporal maxpool k=(3,1,1), st=(s,1,1) → if s>1
BN+ReLU
conv k=(1,1,1), st=(1,1,1),4F
BN
ReLU

(e) SSA-ResNet (Bottleneck)

conv k=(1,1,1), st=(1,1,1), F
BN+ReLU
conv k=(1,3,3), st=(1,s,s), F, group=32
SSA Layer
temporal maxpool k=(3,1,1), st=(s,1,1) → if s>1
BN+ReLU
conv k=(1,1,1), st=(1,1,1),2F
BN
ReLU

(f) SSA-ResNeXt (Bottleneck)

Results on ModelNet10 and ModelNet40

Network	Parameters (Millions)	ModelNet40	ModelNet10
3D ShapeNets	≈38	77%	83.5%
Beam Search	≈0.08	81.26%	88%
3D-GAN	≈11	83.3%	91%
VoxNet	≈0.92	83%	92%
LightNet	≈0.30	86.90%	93.39%
ORION	≈.91	-	**93.8%**
SSA-ResNeXT8 (ours)	≈3.38	**89.5%**	93.3%

Comparisons with the state-of-the-art. The comparison of the SSA-ResNeXT8 with the state-of-the-art methods on the voxelized version of ModelNet40 and ModelNet10 datasets with shapes augmented with 12 orientations. We have considered only volumetric frameworks in this comparison.

Acknowledgements

Gagan Kanojia and Sudhakar Kumawat were supported by TCS Research Fellowships. Shanmuganathan Raman was supported by SERB Core Research Grant and Imprint 2 Grant.

References

1. Hara, K., Kataoka, H. and Satoh, Y., 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. IEEE CVPR (pp. 6546-6555).
2. Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild., CRCV-TR-12-01.