$S^3D$: Single Shot multi-Span Detector via Fully 3D Convolutional Network

Da Zhang$^1$, Xiyang Dai$^2$, Xin Wang$^1$, and Yuan-Fang Wang$^1$

dazhang@cs.ucsb.edu

$^1$UC Santa Barbara & $^2$University of Maryland
Task: Temporal Activity Detection

Input: untrimmed videos

1. Localization: when do activities start/end?
2. Classification: what are the activities?

Detection Results

Pole Vault [228.1 - 236.6s]  Pole Vault [242.0 - 247.7s]
Related Works

Conventional two-stage approach: Proposal + Classification

- **Temporal Proposal**
  - Pole Vault [228.1 - 236.6s]

- **Activity Classifier**
  - Pole Vault [242.0 - 247.7s]

- **Sliding window, DAP, etc.**

- **Two-stream, C3D, etc.**

S-CNN (CVPR 2016), CDC (CVPR 2017), TSN (ICCV 2017), R-C3D (ICCV 2017), SSN (ICCV 2017)
Related Works

Current limitations:

Pole Vault [228.1 - 236.6s]

Ineffective

Activity Classifier

Temporal Proposal

Pole Vault [242.0 - 247.7s]

Inefficient

S-CNN (CVPR 2016), CDC (CVPR 2017), TSN (ICCV 2017), R-C3D (ICCV 2017), SSN (ICCV 2017)
Motivation

Can we do better?

Introducing a novel Single Shot multi-Span Detector ($S^3D$)
Motivation

Quick Summary

- Directly encode entire input video with Conv3D kernels
- Multi-scale default spans associated to temporal feature maps
- End-to-end trainable and single forward-pass inference
Our model takes the whole video stream as input (L frames)
We apply the standard C3D network to extract spatial-temporal features.

We produce a sequence of feature maps that progressively decrease in temporal dimension.
Multi-scale default spans are associated to each temporal feature map
Temporal Feature Layers

Localization and classification results are predicted at each default span.

Loc: \( \Delta(ct, lt) \)
Conf: \((c^1, c^2, \ldots, c^K, c^{act})\)
We apply on top of each feature map a Conv3D filter to produce the results.
**$S^3D$: Convolutional Predictors**

**Temporal Feature Layers**

3D Max pool, Conv3D: $3 \times 1 \times 1 \times 4 \times (K+1+2)$

- **Kernel size**
- **# of scales**
- **Classes + BG** $(c^1, c^2, ..., c^K, c^{act})$
- **Localization offsets** $\Delta(ct, lt)$
Single Shot multi-Span Detector

Training of $S^3$D:

$$Loss = L_{loc}(x, t, g) + \alpha L_{conf}(x, c) + \beta L_{act}(s, c)$$

- **Smooth L1**: $L_{loc}(x, t, g)$
- **Softmax Cross Entropy**: $L_{conf}(x, c)$
- **Sigmoid Cross Entropy**: $L_{act}(s, c)$

**Input Video**
**Base Feature Layers**
**Auxiliary Temporal Feature Layers**
**Temporal Activity Detections**

C3D up to Conv5b layer

Video C3D up to Conv5b layer

3D Max pool, Conv3D: 3x1x1x(4x(K+1+2))

252 Temporal Spans per Video

Temporal NMS
# Quantitative Results

Evaluation: mean Average Precision over 20 activities on THUMOS’14

<table>
<thead>
<tr>
<th>IoU threshold</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-CNN (CVPR 2016)</td>
<td>36.3</td>
<td>28.7</td>
<td>19.0</td>
<td>10.3</td>
<td>5.3</td>
</tr>
<tr>
<td>CDC (CVPR 2017)</td>
<td>40.1</td>
<td>29.4</td>
<td>23.3</td>
<td>13.1</td>
<td>7.9</td>
</tr>
<tr>
<td>SSAD (MM 2017)</td>
<td>43.0</td>
<td>35.0</td>
<td>24.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TCN (ICCV 2017)</td>
<td>-</td>
<td>33.3</td>
<td>25.6</td>
<td>15.9</td>
<td>9.0</td>
</tr>
<tr>
<td>R-C3D (ICCV 2017)</td>
<td>44.8</td>
<td>35.6</td>
<td>28.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SSN (ICCV 2017)</td>
<td><strong>50.6</strong></td>
<td>40.8</td>
<td>29.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SS-TAD (BMVC 2017)</td>
<td>40.1</td>
<td>-</td>
<td>29.2</td>
<td>-</td>
<td>9.6</td>
</tr>
<tr>
<td>S$^3$D (ours)</td>
<td><strong>47.9</strong></td>
<td><strong>41.2</strong></td>
<td><strong>32.6</strong></td>
<td><strong>23.3</strong></td>
<td><strong>14.3</strong></td>
</tr>
</tbody>
</table>

1271 FPS on a single GTX 1080 Ti GPU
Qualitative Results

THUMOS’14 segment: Pole Vault
Qualitative Results

THUMOS’14 segment: Javelin Throw

GT—Background
Pred—Background
Qualitative Results

THUMOS’14 segment: Shotput
Qualitative Results

THUMOS’14 segment: Clean and Jerk
Conclusions

Introduced $S^3D$:

- A novel single-shot end-to-end model for Temporal Activity Detection.
- *Simple*: completely based on Conv3D kernels.
- *Speed*: operates at 1271 FPS on a single GeForce GTX 1080 Ti GPU.

TensorFlow code coming soon at https://github.com/dazhang-cv/S3D
Thank you!

Input Video
Base Feature Layers
Auxiliary Temporal Feature Layers
Temporal Activity Detections