Video Object Segmentation using Deep Learning

Update Presentation, Week 2

Zack While

Advised by: Rui Hou, Dr. Chen Chen, and Dr. Mubarak Shah

May 26, 2017

Youngstown State University
Table of Contents

1. Previous Work
2. Problem Description
3. Current Work
4. Upcoming Work
Previous Work
Accomplished the following:

- Gained familiarity with the programming language Python.
  - Kishan and Vijay led tutorials and answered questions.
  - Worked on programming practice problems.

- Learned the basics of the KERAS deep learning library in Python with the TensorFlow backend.
  - Implemented a classifier for the CIFAR10 data set.
  - Implemented an autoencoder that removed noise from images in the MNIST data set.
Accomplished the following:

- Gained familiarity with the programming language Python.
  - Kishan and Vijay led tutorials and answered questions.
  - Worked on programming practice problems.

- Learned the basics of the KERAS deep learning library in Python with the TensorFlow backend.
  - Implemented a classifier for the CIFAR10 data set.
  - Implemented an autoencoder that removed noise from images in the MNIST data set.
Problem Description
In general, video object segmentation aims to differentiate various parts of a video, normally foreground objects and the background. Regardless of what the object(s) in a video is, we want to be able to classify each pixel of the video as object or background.

Most approaches take a 2-stream approach, with one learning based on appearance and one learning from motion.

This group previously created the T-CNN, an end-to-end 3D CNN network, which takes in video clips directly and was successful in action recognition tasks.

Our goal now is to extend this framework for object segmentation.
In general, video object segmentation aims to differentiate various parts of a video, normally foreground objects and the background.

Regardless of the what the object(s) in a video is, we want to be able to classify each pixel of the video as *object* or *background*.

Most approaches take a 2-stream approach, with one learning based on *appearance* and one learning from *motion*.

This group previously created the T-CNN, an end-to-end 3D CNN network, which takes in video clips directly and was successful in action recognition tasks.

Our goal now is to extend this framework for object segmentation.
In general, video object segmentation aims to differentiate various parts of a video, normally foreground objects and the background.

Regardless of the what the object(s) in a video is, we want to be able to classify each pixel of the video as object or background.

Most approaches take a 2-stream approach, with one learning based on appearance and one learning from motion.

This group previously created the T-CNN, an end-to-end 3D CNN network, which takes in video clips directly and was successful in action recognition tasks.

Our goal now is to extend this framework for object segmentation.
In general, video object segmentation aims to differentiate various parts of a video, normally foreground objects and the background. Regardless of the what the object(s) in a video is, we want to be able to classify each pixel of the video as object or background.

Most approaches take a 2-stream approach, with one learning based on appearance and one learning from motion.

This group previously created the T-CNN, an end-to-end 3D CNN network, which takes in video clips directly and was successful in action recognition tasks.

Our goal now is to extend this framework for object segmentation.
In general, video object segmentation aims to differentiate various parts of a video, normally foreground objects and the background.

Regardless of what the object(s) in a video is, we want to be able to classify each pixel of the video as *object* or *background*.

Most approaches take a 2-stream approach, with one learning based on *appearance* and one learning from *motion*.

This group previously created the T-CNN, an end-to-end 3D CNN network, which takes in video clips directly and was successful in action recognition tasks.

Our goal now is to extend this framework for object segmentation.
Current Work
Goals

1. Become familiar with various 3D CNN architectures, especially those used in object segmentation.

2. Start looking at current Python implementation.
1. Become familiar with various 3D CNN architectures, especially those used in object segmentation.

2. Start looking at current Python implementation.
Read the following papers:

- **Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos** by Hou, Chen, & Shah (arXiv, April 2017)
  - End-to-end deep network
  - Processes motion and spatial information together.
Learning Spatiotemporal Features with 3D Convolutional Networks by Tran et al. (arXiv, October 2015)
- C3D architecture (used in T-CNN paper)
- Found that $3 \times 3 \times 3$ kernel has best performance.
- Captures more temporal information than 2D CNN.

The 2017 DAVIS Challenge on Video Object Segmentation by Pont-Tuset et al. (arXiv, April 2017)
- Multiple annotated objects, occlusions, fast motion, etc.
- Separate ongoing and 2-week challenges.
- Object masks provided in first frame.
- Removed temporal instability from required measures due to increase in occlusions.
Learning Spatiotemporal Features with 3D Convolutional Networks by Tran et al. (arXiv, October 2015)

- C3D architecture (used in T-CNN paper)
- Found that $3 \times 3 \times 3$ kernel has best performance.
- Captures more temporal information than 2D CNN.

The 2017 DAVIS Challenge on Video Object Segmentation by Pont-Tuset et al. (arXiv, April 2017)

- Multiple annotated objects, occlusions, fast motion, etc.
- Separate ongoing and 2-week challenges.
- Object masks provided in first frame.
- Removed temporal instability from required measures due to increase in occlusions.

<table>
<thead>
<tr>
<th></th>
<th>DAVIS 2016</th>
<th>DAVIS 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>train</td>
<td>val</td>
</tr>
<tr>
<td>Number of sequences</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>Number of frames</td>
<td>2079</td>
<td>1376</td>
</tr>
<tr>
<td>Mean number of frames per sequence</td>
<td>69.3</td>
<td>68.8</td>
</tr>
<tr>
<td>Number of objects</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>Mean number of objects per sequence</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos by Jain, Xiong, & Grauman (arXiv, April 2017)

- Uses separate appearance and motion CNN streams.
- Bootstrapped annotated video data from image data sets.
Semantically-Guided Video Object Segmentation by Perazzi et al. (arXiv, December 2016)
- Mainly used when object mask is provided in first frame.
- Semantic prior increases robustness with varying appearance.

Learning Video Object Segmentation from Static Images by Caelles et al. (arXiv, April 2017)
- Current frame’s mask is based on last frame’s mask.
- Uses offline and online training with just image annotations.
Semantically-Guided Video Object Segmentation by Perazzi et al. (arXiv, December 2016)
- Mainly used when object mask is provided in first frame.
- Semantic prior increases robustness with varying appearance.

Learning Video Object Segmentation from Static Images by Caelles et al. (arXiv, April 2017)
- Current frame’s mask is based on last frame’s mask.
- Uses offline and online training with just image annotations.
Learning Video Object Segmentation with Visual Memory by Tokmakov et al. (arXiv, April 2017)

• Currently reading
Received a link to the project’s GitHub page, began reviewing submitted code so far.

Using TensorFlow documentation where needed.
Implementation Details

- Received a link to the project’s GitHub page, began reviewing submitted code so far.

- Using TensorFlow documentation where needed.
Upcoming Work
Plan for Next Week

1. Finishing the literature review (2-4 recommended papers remaining) to generally understand current object segmentation techniques.

2. Continue looking over GitHub repository and asking questions when necessary.

3. Do full TensorFlow tutorial if needed.
Plan for Next Week

1. Finishing the literature review (2-4 recommended papers remaining) to generally understand current object segmentation techniques.

2. Continue looking over GitHub repository and asking questions when necessary.

3. Do full TensorFlow tutorial if needed.
Plan for Next Week

1. Finishing the literature review (2-4 recommended papers remaining) to generally understand current object segmentation techniques.

2. Continue looking over GitHub repository and asking questions when necessary.

3. Do full TensorFlow tutorial if needed.
Plan for Upcoming Weeks


2. Work with research group to decide on a possible architecture for the video object segmentation model.
Plan for Upcoming Weeks


2. Work with research group to decide on a possible architecture for the video object segmentation model.
Thank you!