Dataset Condensation in Videos
Sabrina Lopez, Gaurav Kumar Nayak, Mubarak Shah
University of Central Florida

Project Description
Goal: Develop an algorithm for condensing a large video dataset into a small synthetic video dataset such that the model trained on the small synthetic dataset can obtain similar video classification accuracy compared to the model trained on the full dataset.

Steps:
- Leverage existing video classification-based backbone architectures
- Implement core-set selection techniques (e.g., random) to create baseline algorithm.
- Implement dataset condensation on video dataset via distribution matching (DM) [4].

Dataset
The UCF-101 dataset is an action recognition dataset of realistic action videos, collected from YouTube, comprising of 13,320 videos that are organized into 101 action categories. It provides large variations in camera motion, object appearance and pose, object scale, etc.

Architecture & Method
ResNet-18 3D is a small 3D architecture, which we use as a video backbone architecture to process the video frames. It employs 3D convolutions and obtains a feature representation, which is used for classifying the videos.

Our overall video condensation architecture is shown below:

Differentiable augmentation on both the real and synthetic datasets to:
- exploit information from the real training images more effectively,
- enable the synthetic images to learn certain prior knowledge in the real images, and
- use the synthetic images with various data augmentation strategies to train different deep neural network architectures.

We used random horizontal and vertical flips as our augmentations.

The concept of distribution matching (DM) loss and its advantage is shown below:

For experiments, the MC3-18 and ResNet-34 3D architectures will be used to make comparisons of results between different architectures and architecture depths.

MC3-18 is a mixed convolutional architecture consisting of 18 convolutional layers like ResNet-18 3D.

ResNet-34 3D is the same as the ResNet-18 3D architecture but with 34 instead of 18 convolutional layers.

We conducted the baseline experiments with the algorithm based on the core-set selection to find the results by changing different architectures and architecture depths.

Leverage existing core-set selection technique shows an increase in performance as more data is added to the dataset. This is promising for future endeavors into this novel problem.

Experiments
We intended to complete distribution matching component of the algorithm, implement the same experiments to the HMDB51 dataset and apply distribution matching on videos with image-based 2D backbone, specifically using superimages [5], and convert video condensation to superimage condensation.

Quantitative Results

<table>
<thead>
<tr>
<th>Random % of Data on ResNet-18</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1 Acc</td>
<td>24.757</td>
<td>35.141</td>
<td>46.496</td>
<td>64.297</td>
<td>81.228</td>
</tr>
<tr>
<td>Top 5 Acc</td>
<td>48.389</td>
<td>59.488</td>
<td>70.23</td>
<td>82.762</td>
<td>97.903</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random % of Data on MC3-18</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1 Acc</td>
<td>28.747</td>
<td>40.409</td>
<td>41.228</td>
<td>42.251</td>
<td>73.862</td>
</tr>
<tr>
<td>Top 5 Acc</td>
<td>49.719</td>
<td>61.893</td>
<td>72.481</td>
<td>74.322</td>
<td>96.675</td>
</tr>
</tbody>
</table>

Conclusion
We proposed using core-set techniques and distribution matching as means to solve dataset condensation in videos. While the latter is still in development, our baseline performance for the random core-set selection technique shows an increase in performance as more data is added to the dataset. This is promising for future endeavors into this novel problem.

Future Work
We intend to complete distribution matching component of the algorithm, implement the same experiments to the HMDB51 dataset and apply distribution matching on videos with image-based 2D backbone, specifically using superimages [5], and convert video condensation to superimage condensation.

References