Shuffle and Learn

Donsuk Lee
Rui Shu
Representation learning for videos

- Rich in information
- Spatiotemporal reasoning necessary for many cognitive tasks
- Representations useful for many computer vision tasks
Representation learning for videos

- Rich in information
- Spatiotemporal reasoning necessary for many cognitive tasks
- Representations useful for many computer vision tasks

(image credit: github.com/dyelax)
Previous Work
Slow Features

- Temporally close images: extracted features should be close
- Temporally far images: extracted features should be far

(Wiskott, 2002)
Slow Features

- Temporally close images: extracted features should be close
- Temporally far images: extracted features should be far

(Goroshin, 2015)
Slow Features

- Temporally close images: extracted features should be close
- Temporally far images: extracted features should be far

(Jayaraman, 2015)
Prediction-Based

- Predict input frames from latent variables (reconstruction)
- Predict next frames (future prediction)

(Srivastava, 2015)
Prediction-Based: Limitations

- Video prediction: extremely high-dimensional task
- Difficult to train
- Difficult to scale to large video datasets

(Srivastava, 2015)
A New Approach
Verification-Based: Temporal Order Verification

- Is the frame order correct or incorrect (shuffled)?
- Simpler task than video prediction
- Requires spatiotemporal reasoning
Contribution: Tuple Verification

- A verification-based unsupervised learning from videos
- Tuple verification:
  - Learns representations that capture spatiotemporal information
  - Learns representations that transfer to downstream video tasks
  - Compliments representations learned from supervised vision tasks
Tuple Verification: Formulation

- Sample \((f_a, f_b, f_c, f_d, f_e)\) from high-motion temporal windows
- Subsample 3 frames:
  - "Correct" sequence: temporal forward or reverse direction
  - "Incorrect" sequence: temporally out of order
Tuple Verification: Architecture

Input Tuple

AlexNet architecture

fc8

classification

Shared parameters
Tuple Verification: Technical Considerations

- How many frames needed for tuple verification?
- How big should the temporal window be?
- Should training set class ratio be balanced or imbalanced?
Empirical Ablation Analysis: #Frames per tuple

- How many frames are needed for correct ordering?
  - Two Frames
  - Three Frames

Ambiguous

Using more (4,5) frames per tuple did not show significant improvement

(Image credit: https://youtu.be/sRdWR3ONPJI)
Empirical Ablation Analysis: Sampling

- Large $\tau_{\text{max}}$ makes positive samples difficult
- Small $\tau_{\text{min}}$ makes negative samples difficult

(a) Varying temporal sampling

<table>
<thead>
<tr>
<th>$\tau_{\text{max}}$</th>
<th>$\tau_{\text{min}}$</th>
<th>Tuple Pred.</th>
<th>Action Recog.</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>15</td>
<td>60.2</td>
<td>47.2</td>
</tr>
<tr>
<td>60</td>
<td>15</td>
<td>72.1</td>
<td>50.9</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>64.3</td>
<td>49.1</td>
</tr>
</tbody>
</table>

$\tau_{\text{max}} = |b - d|$

$\tau_{\text{min}} = \min(|a - b|, |d - e|)$

→ Better results with more difficult samples
Empirical Ablation Analysis: Class Ratios

- Varying ratio between positive and negative samples in each mini-batch during training phase

<table>
<thead>
<tr>
<th>Class Ratio</th>
<th>Tuple Pred.</th>
<th>Action Recog.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neg</td>
<td>Pos</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>52.1</td>
</tr>
<tr>
<td>0.65</td>
<td>0.35</td>
<td>68.5</td>
</tr>
<tr>
<td>0.75</td>
<td>0.25</td>
<td><strong>72.1</strong></td>
</tr>
<tr>
<td>0.85</td>
<td>0.15</td>
<td>67.7</td>
</tr>
</tbody>
</table>
Representation Analysis
Evaluation of the Representation

1. What does the representation learned on tuple verification task capture?
2. Evaluation on Action Recognition task
3. Evaluation on Pose Estimation task
What does the tuple verification task capture?

Nearest Neighbor retrieval

- **ImageNet** pre-trained network focuses on *scene semantics*

- **Tuple verification** pre-trained network focuses on *human pose*
What does the tuple verification task capture?

Visualizing pool5 responses

- Each row are the top image regions for a unit from the pool5 layer
- Many units show preference for human body parts and pose
Action-Recognition
Action Recognition Task

Unsupervised Tuple Verification

Supervised Action Recognition

UCF 101

Figure 1: Two-stream architecture for video classification.

(Simoyan et al.)
Comparison on Action Recognition

<table>
<thead>
<tr>
<th>Pre-training Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
</tr>
<tr>
<td>Two Close, Two Order, Three Order, DrLim, TempCoh, Obj.Patch</td>
</tr>
<tr>
<td>Supervised</td>
</tr>
<tr>
<td>UCF Sup., ImageNet, ImageNet+UCF Sup</td>
</tr>
<tr>
<td>Supervised + Unsupervised</td>
</tr>
<tr>
<td>ImageNet+Tuple verification</td>
</tr>
</tbody>
</table>

Figure 1: Two-stream architecture for video classification.

(Simoyan et al.)
Comparison on Action Recognition

- “Random initialization” v.s. “Tuple verification Pretraining”

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Initialization</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF101</td>
<td>Random</td>
<td>38.6</td>
</tr>
<tr>
<td></td>
<td>(Ours) Tuple verification</td>
<td><strong>50.2</strong></td>
</tr>
<tr>
<td>HMDB51</td>
<td>Random</td>
<td>13.3</td>
</tr>
<tr>
<td></td>
<td>UCF Supervised</td>
<td>15.2</td>
</tr>
<tr>
<td></td>
<td>(Ours) Tuple verification</td>
<td><strong>18.1</strong></td>
</tr>
</tbody>
</table>

→ Demonstrates informativeness of “Tuple verification pretraining"
Comparison on Action Recognition

- Comparison with other unsupervised baselines

<table>
<thead>
<tr>
<th>Unsup Method</th>
<th>Two Close</th>
<th>Two Order</th>
<th>DrLim [40]</th>
<th>TempCoh [38]</th>
<th>Three Order (Ours)</th>
<th>Obj. Patch* [52]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc. UCF101</td>
<td>42.3</td>
<td>44.1</td>
<td>45.7</td>
<td>45.4</td>
<td>50.9</td>
<td>40.7</td>
</tr>
<tr>
<td>Acc. HMDB51</td>
<td>15.0</td>
<td>16.4</td>
<td>16.3</td>
<td>15.9</td>
<td>19.8</td>
<td>15.6</td>
</tr>
</tbody>
</table>

→ “Three Order” more informative than others
Comparison on Action Recognition

- Combining unsupervised and supervised training

<table>
<thead>
<tr>
<th>Initialization</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>13.3</td>
</tr>
<tr>
<td>(Ours) Tuple verification</td>
<td>18.1</td>
</tr>
<tr>
<td>UCF sup.</td>
<td>15.2</td>
</tr>
<tr>
<td>ImageNet</td>
<td>28.5</td>
</tr>
<tr>
<td>(Ours) ImageNet + Tuple verification</td>
<td>29.9</td>
</tr>
<tr>
<td>ImageNet + UCF sup.</td>
<td>30.6</td>
</tr>
</tbody>
</table>

“ImageNet+Tuple verification” better than “ImageNet” and “Tuple verification”

→ “Tuple verification” adds complementary information to “ImageNet”
Pose Estimation
Comparison on Pose Estimation

### Pre-training Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>Tuple Verif., DrLim, Obj.Patch</td>
</tr>
<tr>
<td>Supervised</td>
<td>UCF Sup., ImageNet</td>
</tr>
<tr>
<td>Supervised + Unsupervised</td>
<td>ImageNet+Tuple Verif.</td>
</tr>
</tbody>
</table>

#### Pre-training Stage
- **UCF 101**
  - **FLIC or MPII**
  - **Deeppose** *(Toshev et al.)*

#### Fine-Tuning & Testing
- Initial stage
- Stage s
- DNN-based regressor
- Send refined values to next stage
Comparison on Pose Estimation

<table>
<thead>
<tr>
<th>Init.</th>
<th>PCK for FLIC</th>
<th>PCKh@0.5 for MPII</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>wr</td>
<td>elb</td>
</tr>
<tr>
<td>Random Init.</td>
<td>53.0</td>
<td>75.2</td>
</tr>
<tr>
<td>Tuple Verif.</td>
<td><strong>69.6</strong></td>
<td><strong>85.5</strong></td>
</tr>
<tr>
<td>Obj. Patch[52]</td>
<td>58.2</td>
<td>77.8</td>
</tr>
<tr>
<td>DrLim[40]</td>
<td>37.8</td>
<td>68.4</td>
</tr>
<tr>
<td>UCF Sup.</td>
<td>61.0</td>
<td>78.8</td>
</tr>
<tr>
<td>ImageNet</td>
<td>69.6</td>
<td>86.7</td>
</tr>
<tr>
<td>ImageNet + Tuple</td>
<td><strong>69.7</strong></td>
<td><strong>87.1</strong></td>
</tr>
</tbody>
</table>

- Outperforms other unsupervised methods & UCF Sup.
- “ImageNet+Tuple” better than both “ImageNet” and “Tuple Verif.”
  → “Tuple Verif.” complementary to “ImageNet”
A verification-based unsupervised learning from videos

- Alternative: a simpler prediction-based approach, such as temporal jigsaw solving?
- Scalability: does the approach actually work for higher-dimensional video datasets?

Tuple verification:

- Learns representations that capture spatiotemporal information
  - Nearest neighbor retrieval: Representations capture semantically meaningful distance for pose
  - Pool5 activations: Strong focus on human body and pose
- Learns representations that transfer to downstream video tasks
  - Tuple verification initialization performs better than random initialization
  - Would be interesting to consider multi-task formulation
- Compliments representations learned from supervised vision tasks
  - ImageNet initialization improved by Tuple verification fine-tuning
  - Might Tuple verification overwrite some of ImageNet's useful representations?