Relational Action Forecasting

Chen Sun; Abhinav Shrivastava et al., CVPR 2019

Presenters: Wanda Wang, Vincent Chung, Shreyas Hosamane
Problem Statement

Given a history of $H$ previous frames, the goal is to detect actors and to predict their future actions for the next $T$ frames.
Why is action forecasting important?

Self-driving cars

Human interaction robot
Related work

• Action recognition
• Future prediction
• Relational reasoning
Action recognition

• Action classification

• Temporal action localization

• Spatio-temporal action detection
Action recognition: Action classification

\[ p(a^{0:T} | V^{0:T}) \]

\[ V^{0:T} : \text{frames of the entire video} \]

\[ a^{0:T} : \text{predicted action labels} \]
Action recognition: Temporal action localization

\[ p(\mathbf{a}^{0:T}, \mathbf{b}^{0:T} | \mathbf{V}^{0:T}) \]

\( \mathbf{V}^{0:T} \): frames of the entire video
\( \mathbf{a}^{0:T} \): predicted action labels
\( \mathbf{b}^{0:T} \): predicted locations of actions
Action recognition: Spatio-temporal action detection

\[ p(a_{1:N}^{0:T}, b_{1:N}^{0:T} | V^{0:T}) \]

- \( N \): number of predicted actors
- \( V^{0:T} \): frames of the entire video
- \( a_{1:N}^{0:T} \): predicted action labels of \( N \) actors
- \( b_{1:N}^{0:T} \): predicted locations of \( N \) actors

Future prediction

Pixel prediction

Trajectory prediction

CVPR 2019 Oral Session 1-1C: Action & Video : https://www.youtube.com/watch?v=JwaBi_2JFeU&t=1105s
Relational reasoning

- This work aims to capture human-human relationships to reason about future actions
Set of actions for last observed frame

- Sitting
- Talking
- Holding
- Standing
- Hugging
- Standing

CVPR 2019 Oral Session 1-1C: Action & Video : https://www.youtube.com/watch?v=JwaBi_2JFeU&t=1105s
Action forecasting

Sitting
Talking

Next:
Clink glass

Holding
Standing

Next:
Serving

Hugging
Standing

Next:
Kissing

CVPR 2019 Oral Session 1-1C: Action & Video : https://www.youtube.com/watch?v=JwaBi_2JFeU&t=1105s
Proposed Approach

Mathematical representation

\[ p(N, b_{1:N}^0, a_{1:N}^{0:T} | V^{-H:0}) \]

\( V^{-H:0} \rightarrow \) visual history of H previous frames

\( N \rightarrow \) number of predicted actors

\( b_{1:N}^0 \rightarrow \) predicted locations (bounding boxes) of N actors at time \( t = 0 \)

\( a_{1:N}^{0:T} \rightarrow \) predicted action labels for N actors for time \( t = 0:T \)
Proposed Approach

Creating the nodes in the graph

\[
p(N, b_{1:N}^0, a_{1:N}^{0:T}|V^{-H:0}) = \delta(N, b_{1:N}^0 | f_{RPN}(V^0)) p(a_{1:N}^0 | h_{1:N}^0, b_{1:N}^0, V^{-H:0}) \prod_{t=1}^{T} p(a_{1:N}^t | h_{1:N}^t, a_{1:N}^{t-1}, h_{1:N}^{t-1})
\]

\[
\prod_{n=1}^{N} \text{Cat}(a_n^0 | f_{CLS}(h_n^0)) \delta(h_n^0 | f_{ROI}(f_{S3D}(V^{-H:0}), b_n^0))
\]
Proposed Approach

Modeling node dynamics
Proposed Approach

Modeling the edges

\[
p(a_{1:N}^t, h_{1:N}^t | a_{1:N}^{t-1}, h_{1:N}^{t-1}) = \prod_{n=1}^{N} \text{Cat}(a_n^t | f_{CLS}(h_n^t)) \delta(h_n^t | f_{RNN}(h_{n-1}^{t-1}, a_n^{t-1})) \delta(h_{n-1}^{t-1} | f_{GNN}(h_{1:N}^{t-1}))
\]
Proposed Approach

Modeling the edges

\[
e_{ij} = f_{\text{edge}}(h_i, h_j)
\]

\[
\tilde{h}_i = f_{\text{node}} \left( \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} e_{ij} \right)
\]

\[
h_i^t = f_{\text{RNN}}(\tilde{h}_i^{t-1}, a_i^{t-1})
\]

sensitive to noisy nodes
Proposed Approach

Modeling the edges

\[ z_i = \sum_j \alpha_{ij} h_j \]

\[ \alpha_{ij} = \text{softmax}(f_{\text{attn}}(e_{ij})) \]

\[ \tilde{h}_i = f_{\text{node}}(z_i) \]

\[ \tilde{h}_i = f_{\text{node}}([h_i; z_i]) \]

difficulty distinguishing node features from neighbor features

\[ \tilde{h}_{1:N} = f_{\text{GNN}}(h_{1:N}) = \prod_{i=1}^N \delta(\tilde{h}_i | f_{\text{node}}([h_i, z_i])) \delta(z_i) \sum_j \alpha_{ij} h_j \prod_{j=1}^N \delta(\alpha_{ij} | S_j(f_{\text{attn}}(e_{i,1:N})) \delta(e_{ij} | f_{\text{edge}}(h_i, h_j))) \]
Proposed Approach

Overall DR^2N model
Proposed Approach

Training

\[ L^{\text{total}} = \alpha L^{\text{loc}} + \sum_{t=0}^{T} \beta_t L^{\text{cls}}_t \]

- \( L^{\text{loc}} \rightarrow \) bounding box localization loss
- \( L^{\text{cls}}_t \rightarrow \) action classification loss at time \( t \)
- \( \alpha, \beta_t \rightarrow \) scalar weights

\[ \alpha = 1 \]
\[ \beta_0 = 1, \text{ linearly decrease such that } \beta_t = 0.5 \]
Proposed Approach

Implementation details

• RPN: ResNet-50 initialized with pre-trained ImageNet weights
• Feature network: S3D-G weights pre-trained from Kinetics-400

• GRU: RNN architecture to model action dynamics

• Synchronous SGD with batch size 4 per GPU
• 10 GPUs in total

• Warm-start, cosine learning rate decay
• Gradient multiplier
Experimental setup

AVA (Atomic Visual Actions)
- Large-scale action detection dataset (430 15-min clips)
- 80 atomic visual actions
- Multiple actors with multiple action labels
- If IoU > 0.5 and action label is correct → true positive

Orange: 1 pose action
Red: 0-3 interactions with objects
Blue: 0-3 interactions with other people

Atomic Actions

The text shows pairs of atomic actions for the people in red bounding boxes.

J-HMDB (Joint-annotated Human Motion Database)

- 21 categories (1.4 seconds/clip)
- One actor with a single action label
- Test for an early action classification problem
Methods for ablation study

Single-head
- Train a separate model for each future time step

Multi-head
- Train one model for all future t

GRU
- Predicted from hidden states of GRU
- Without edges (relations)

Graph-GRU
- With edges
  - **RN**: equal weights
  - **GAT**: weighted differently but is not sure if it should focus on features from itself or features from neighbors
  - **DR²N**: proposed method
**Action forecasting results on AVA dataset**

<table>
<thead>
<tr>
<th>Method</th>
<th>Dynamics Model</th>
<th>Relation Model</th>
<th>$t = 0$</th>
<th>$t = 1$</th>
<th>$t = 2$</th>
<th>$t = 3$</th>
<th>$t = 4$</th>
<th>$t = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-head</td>
<td>-</td>
<td>-</td>
<td>19.1</td>
<td>7.8</td>
<td>5.3</td>
<td>4.2</td>
<td>2.6</td>
<td>1.8</td>
</tr>
<tr>
<td>Multi-head</td>
<td>-</td>
<td>-</td>
<td>16.0</td>
<td>9.4</td>
<td>6.8</td>
<td>5.4</td>
<td>4.3</td>
<td>3.6</td>
</tr>
<tr>
<td>GRU</td>
<td>GRU</td>
<td>-</td>
<td>18.7</td>
<td>13.1</td>
<td>10.3</td>
<td>8.0</td>
<td>6.7</td>
<td>5.7</td>
</tr>
<tr>
<td>Graph-GRU</td>
<td>GRU</td>
<td>RN [51]</td>
<td>17.3</td>
<td>12.3</td>
<td>9.9</td>
<td>7.7</td>
<td>6.5</td>
<td>5.3</td>
</tr>
<tr>
<td>Graph-GRU</td>
<td>GRU</td>
<td>GAT [61]</td>
<td>16.4</td>
<td>12.3</td>
<td>9.3</td>
<td>7.3</td>
<td>6.2</td>
<td>5.2</td>
</tr>
<tr>
<td>Graph-GRU</td>
<td>GRU</td>
<td>DR$^2$N (Us)</td>
<td>20.4</td>
<td>14.4</td>
<td>11.2</td>
<td>9.3</td>
<td>7.5</td>
<td>6.8</td>
</tr>
</tbody>
</table>

4. DR$^2$N outperforms the other two methods
Short duration action is hard to predict

Change in AP performance from $t = 0$ to $t = 1$
Most gains for actions with explicit interactions or where other actors provide useful context

Change in AP performance from adding graph connections at t = 0
Visualization of discriminative relations

- Orange boxes: query actor (whose actions are to be predicted)
- Blue boxes: top 3 actor proposals with highest attention weights
Results Visualization

Prediction: Eat
Ground truth: Eat

Prediction: Get Up
Ground truth: Get Up

CVPR 2019 Oral Session 1-1C: Action & Video : https://www.youtube.com/watch?v=JwaBi_2JFeU&t=1105s
Results Visualization

Prediction: Put down
Ground truth: Put down

Prediction: Smoke
Ground truth: Smoke

CVPR 2019 Oral Session 1-1C: Action & Video : https://www.youtube.com/watch?v=JwaBi_2JFeU&t=1105s
Results Visualization

Prediction: Read
Ground truth: Read

Prediction: Open (door)
Ground truth: Open (door)
Results Visualization (Failure modes)

Prediction: Get up  
Ground truth: Keep kneeling  
Reason: Wrong duration

Prediction: Close  
Ground truth: Open  
Reason: Multiple futures

CVPR 2019 Oral Session 1-1C: Action & Video : https://www.youtube.com/watch?v=JwaBi_2JFeU&t=1105s
Early action classification on J-HMDB

• Take the first K% of frames and predict the label of the clip

<table>
<thead>
<tr>
<th>Model</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soomro et al. [55]</td>
<td>≈ 5</td>
<td>≈ 12</td>
<td>≈ 21</td>
<td>≈ 25</td>
<td>≈ 30</td>
</tr>
<tr>
<td>Singh et al. [54]</td>
<td>≈ 48</td>
<td>≈ 59</td>
<td>≈ 62</td>
<td>≈ 66</td>
<td>≈ 66</td>
</tr>
<tr>
<td>GRU</td>
<td>52.5</td>
<td>56.2</td>
<td>61.1</td>
<td>65.2</td>
<td>65.9</td>
</tr>
<tr>
<td>GAT [61]</td>
<td>58.1</td>
<td>61.8</td>
<td>64.4</td>
<td>68.7</td>
<td>68.8</td>
</tr>
<tr>
<td>DR²N</td>
<td>60.6</td>
<td>65.8</td>
<td>68.1</td>
<td>71.4</td>
<td>71.8</td>
</tr>
</tbody>
</table>
Strengths & Weaknesses

Strengths

• Proposed method outperforms existing methods on a complex dataset like AVA
• DR$^2$N also improves performance on task of early action classification from previous SOTA of 48% to 60%

Weaknesses

• In the relations visualization, duplicate detections exist that might unnecessarily increase graph size and also affect the weights
• Struggles with actions having multiple possible futures and when duration is hard to predict
Possible extensions

• Implement NMS to suppress duplicate detections and obtain better graph representations
• Conduct study with different values of H to demonstrate effect of temporal context duration
• Model human-object relationships in addition to human-human
Conclusion

• This paper proposed a multi-person action forecasting model (DR$^2$N) that considers temporal and spatial interactions among actors
• AVA and J-HMDB datasets used for evaluation demonstrate effectiveness