Learning to act by predicting the future

Alexey Dosovitskiy and Vladlen Koltun
Intel Labs, Santa Clara

ICLR 2017, Toulon, France
Sensorimotor control

Aim: produce useful motor actions based on sensory inputs
• Aim: maximize the (discounted) sum of future rewards
Training signal

• Standard RL: scalar reward

• Reality: rich sensory stream
Goals

- Standard RL: a single goal given by the reward

- Reality: a wide variety of potential goals
Training procedure

• Standard RL: maximize future discounted reward

• Reality: learn about the world

Image from https://www.flickr.com/photos/lupuca/8720602322
Recent related work

- **Training signal**
  - Jaderberg et al., ICLR 2017

- **Goals**
  - Schaul et al., ICML 2015

- **Training procedure**
  - Exploration & intrinsic motivation
Direct future prediction

• Control as “future-supervised” learning

• Instead of learning to maximize returns, learn to predict the future

How to represent the future?
Naïve approach: predict pixels

• Simply predict the future observation
  - Oh et al. 2015, Finn et al. 2016, Chiappa et al. 2017, ...

• Problem: uncertainty!

• We only need to predict relevant values
Our approach: predict measurements

• Predict the future values of measurements available to the agent

• Assumption: goals (objective functions) can be expressed as functions of measurements
Using future predictions to act

Objective is linear in future measurements
Direct future prediction

• Predict the future measurements for each action
• Simple supervised learning

- The future is stochastic
- ... and depends on the future actions
Network architecture

- **Image** \( S(s) \)
- **Measurements** \( M(m) \)
- **Goal** \( G(g) \)
- **Action** \( A(j) \)
- **Expectation** \( E(j) \)
- **Prediction** \( f \)
- **Target**

Can change dynamically!
Technical details

• One frame as input, no memory
• Predict multiple future steps: 1, 2, 4, 8, 16, 32
• Epsilon-greedy policy
• Small experience replay
• Parallel exploration – 8 copies of the agent
Experiments: ViZDoom environment

- Based on Doom
- Natural measurements: ammo, health, frags
- Wide variety of tasks and scenarios
ViZDoom – tasks

D1: Basic

D2: Navigation

D3: Battle

D4: Battle 2
Comparison to existing methods

D1: Basic

D2: Navigation

D3: Battle

D4: Battle 2
Learning to Act by Predicting the Future

Alexey Dosovitskiy    Vladlen Koltun
Generalization

This was testing on the training set

Can we generalize?
Generalization across goals

• Train with a random goal vector in every episode
  - Uniform [0,1]
  - Uniform [-1,1]

• Change the goal vector at test time
  - The end goal does not have to be known at training time!
• Goal-agnostic training performs very close to training with a fixed goal
• Generalizes to different goals much better
• Train with randomized textures
Generalization across environments

**Battle - #frags**

- **Train - fixed textures**: 35
- **Test - seen textures**: 5
- **Test - new textures**: 0

Legend:
- Blue: Test - seen textures
- Orange: Test - new textures
Generalization across environments

- Good generalization to previously unseen textures and labyrinth layouts
# ViZDoom Competition: Full Deathmatch

<table>
<thead>
<tr>
<th>Place</th>
<th>Team</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IntelAct</td>
<td>29</td>
<td>21</td>
<td>23</td>
<td>21</td>
<td>6</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>30</td>
<td>32</td>
<td>33</td>
<td>35</td>
<td>256</td>
</tr>
<tr>
<td>2</td>
<td>The Terminators</td>
<td>22</td>
<td>17</td>
<td>21</td>
<td>15</td>
<td>13</td>
<td>12</td>
<td>6</td>
<td>5</td>
<td>14</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>164</td>
</tr>
<tr>
<td>3</td>
<td>TUHO</td>
<td>8</td>
<td>11</td>
<td>13</td>
<td>12</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-4</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>51</td>
</tr>
<tr>
<td>4</td>
<td>ColbyCS</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>5vision</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>Ivomi</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>-4</td>
<td>-4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>7</td>
<td>PotatoesArePrettyOk</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-9</td>
</tr>
</tbody>
</table>
Summary

• Simple finite-horizon supervised training performs very well on visuomotor control tasks

• Predicting measurements instead of rewards:
  - Better training signal
  - Flexible goal setting, goal-agnostic learning

• Training with random textures leads to good generalization across environments
The end

More details in our poster yesterday!

Code: https://github.com/IntelVCL/DirectFuturePrediction
Includes our environments and pre-trained models

Videos: IntelVCL youtube channel