Aristotle University of Thessaloniki
Department of Informatics

# USING PART-BASED REPRESENTATION FOR EXPLAINABLE DEEP REINFORCEMENT LEARNING

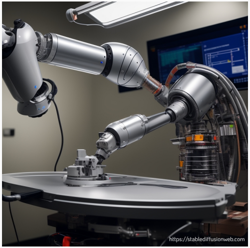
M. Kirtas, K. Tsampazis, L. Avramelou, N. Passalis, A. Tefas
{eakirtas, tsampaka, avramell, passalis, tefas}@csd.auth.gr

# PRESENTATION OUTLINE

# INTRODUCTION

# CHALLENGES IN DRL MODEL-BASED EXPLANATION

- **The use of DRL agents in critical environments**, where safety is highly prioritized, is hindered due to the limited transparency of the models.
- **Extracting the rationale** of a DL model in a human-interpretable way **remain a challenging task**.




1

[1]The images are generated with a Stable Diffusion model

The ability of doing human interpretable models would allow us to:

- Improve the trustworthiness of the model
- Prevent failures
- Improve performance
- Augment human collaboration and users experience

Extracting a part-based representation of DL models provides a great potential to design inherently explainable models, providing transparent mechanism to decision-making process.

- **Canceling neurons are eliminated.**
- Their representation is based on simple **addition of latent causes** acquired from feature representation.
- Hierarchical representation of data, where **higher-level parts are composed of lower-level parts.**
- Part-based representations align more **closely with human intuition.**
- Better **visualizations** allowing model interpretation.

Part-based learning is conceptually tied to human cognition[2]

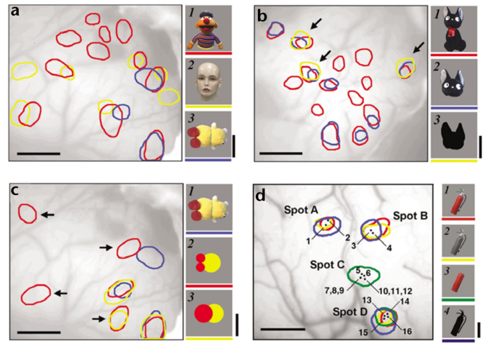



**Figure:** Representation of complex object images and simplification of them in area TE (Source 2)

[2] Tsunoda, K., Yamane, Y., Nishizaki, M. et al. Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. Nat Neurosci 4, 832–838 (2001)

Training part-based learning includes:

- Sign constraints to model's parameters, leading to training difficulties, such as **instabilities and convergence issues**
- **Different initialization and optimization schemes.**

## CHALLENGES IN PART-BASED REPRESENTATION

Training part-based learning includes:

- Sign constraints to model's parameters, leading to training difficulties, such as **instabilities and convergence issues**
- **Different initialization and optimization schemes.**

Existing approaches for part-based learning are limited:

- Applied solely on autoencoders, and **models that are not usually used in DRL**.
- Resulting in a significant **performance degradation**.
- **Making them unsuitable for RL.**

# PROPOSED METHOD

We propose a training approach for actor models in RL approaches, allowing for extracting part-based representations that can provide increased interpretability.

The proposed method includes:

1. An exponential distribution-based **positive-only initialization scheme** for actor model.
2. An alternative **sign-preserving optimization method** to Stochastic Gradient Ascent (SGA), allows one to train the actor model in a non-negative manner.

The proposed pipeline enables more efficient training of inherently explainable models that are based on the non-negative part-based representation of the actor.

**PPO utilizes actor-critic networks**, where the actor parameters are denoted as $\boldsymbol{\theta}$ and critic ones as $\tilde{\boldsymbol{\theta}}$. The PPO method **trains the actor based on the policy gradient** approach, while the **critic evaluates the actions by computing the corresponding state/action values.**

The objective function of the actor is defined as:

$$L^{\mathrm{actor}}(s_t; \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \mathbb{E}_t \left[ \min \left( r_t^{\mathrm{clip}}(\boldsymbol{\theta}) A_t(\tilde{\boldsymbol{\theta}}), r_t^{\mathrm{clip}}(\boldsymbol{\theta}) A_t(\tilde{\boldsymbol{\theta}}) \right) \right] \in \mathbb{R}, \qquad (1)$$

where $A_t(\tilde{\boldsymbol{\theta}})$ is the advantage and $r_t^{\mathrm{clip}}(\boldsymbol{\theta})$ the clipped policy ratio between policy parameterization.

To this end, the Temporal Difference (TD) residual for each time step t is calculated as:

$$\delta_t(\tilde{\boldsymbol{\theta}}) = R_t + \gamma V^{\pi}_{\tilde{\boldsymbol{\theta}}_t}(s_{t+1}) - V^{\pi}_{\tilde{\boldsymbol{\theta}}_t}(s_t) \in \mathbb{R}, \tag{2}$$

where $R_t$ is the reward the agent receives at time step t, $V^{\pi}_{\tilde{\boldsymbol{\theta}}_t}(s_t)$ is the value estimation predicted by the critic policy $\pi$ for current state $s_t$ based on critic parameter $\tilde{\boldsymbol{\theta}}_t$, $\gamma$ is the discount factor and $\lambda$ is the smoothing parameter. In this work, we use $\gamma = 0.99$ and $\lambda = 0.95$.

Then, the advantage $A_t$ is defined as:

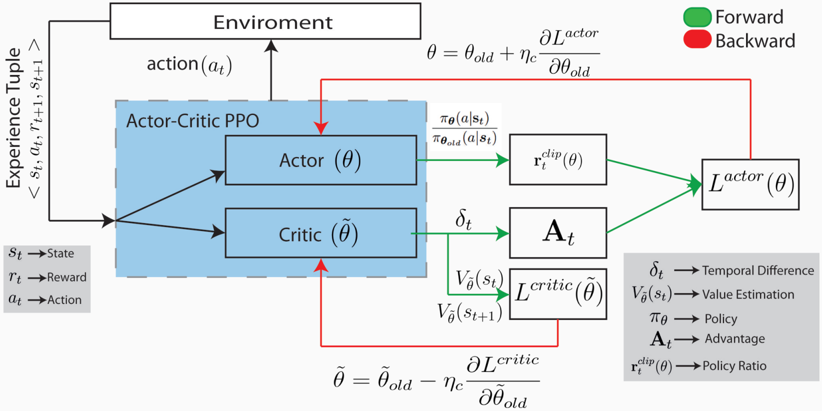
$$A_t(\tilde{\boldsymbol{\theta}}) = \sum_{i=0}^{n-t} \gamma^i \lambda^i \delta_{t+i}(\tilde{\boldsymbol{\theta}}) \in \mathbb{R}, \tag{3}$$

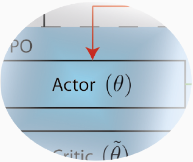
where n is the total number of steps within an episode and t is the time step.

On the other hand, the critic network is typically trained to minimize the temporal difference between the returns and it is formulated as:

$$L^{critic} = \mathbb{E}_t[\delta_t(\tilde{\boldsymbol{\theta}})^2] \in \mathbb{R} \tag{4}$$

# PROXIMAL POLICY OPTIMIZATION

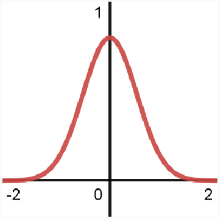





**Traditionally Used**

$$\theta \sim (0, \sigma) \in \mathbb{R}$$

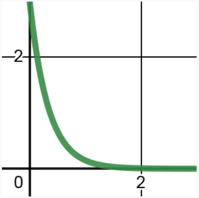
Depends on initilization scheme



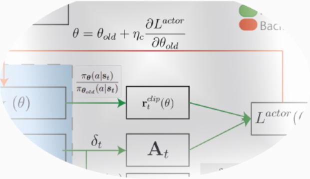

**Proposed**

$$\theta \sim \mathrm{Exp}(\lambda) = \frac{\ln(U(0,1))}{\lambda} \in \mathbb{R}_+$$

Hyperparameter (default λ=100)

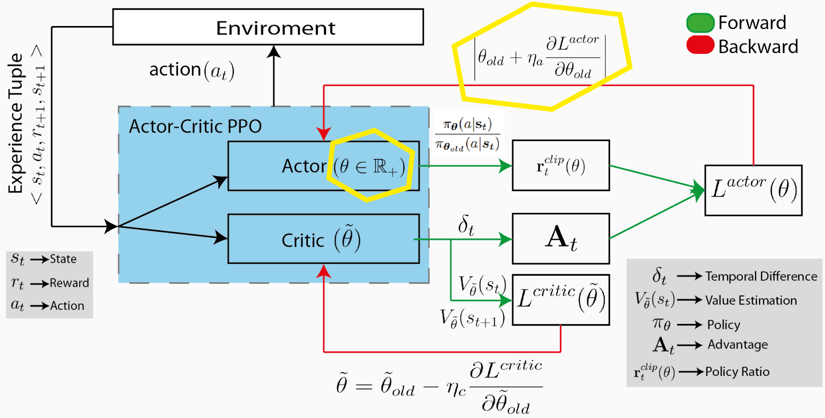


# PROPOSED OPTIMIZATION OF THE ACTOR



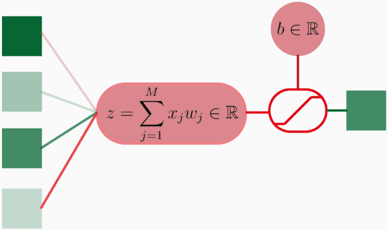

| Traditionally Used | Proposed |
|---|---|
| $\theta = \theta_{old} + \eta_a \frac{\partial L^{actor}}{\partial \theta_{old}} \in \mathbb{R}$ | $\theta = \left\| \theta_{old} + \eta_a \frac{\partial L^{actor}}{\partial \theta_{old}} \right\| \in \mathbb{R}_+$ |
| * Optimizes the actor without constraining the sing of parameters<br>* Do not results in part-based representation | * Preserves the initial sign of the parameter<br>* Eliminates canceling neurons<br>* Achieves part-based representation |

# PROPOSED METHOD



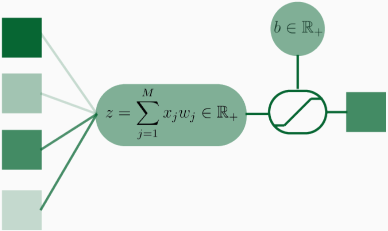
# PART-BASED NEURON

## Typical Neuron

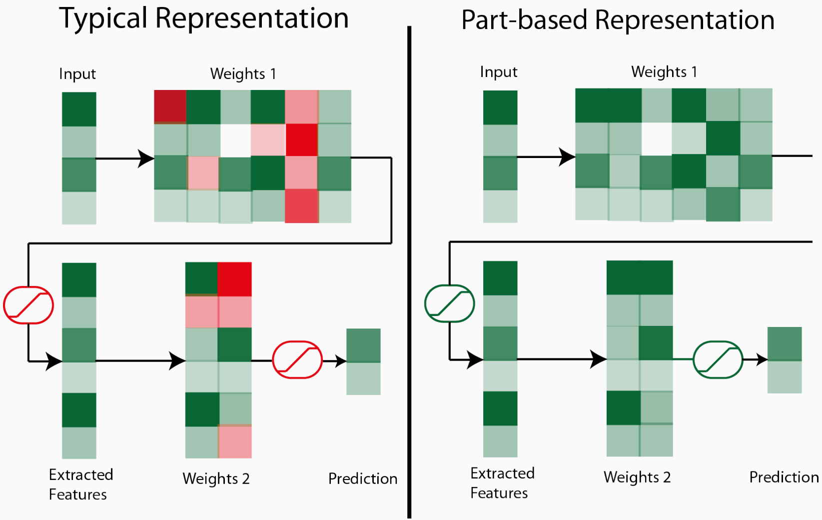



(-) Include both excitatory and inhibitory synapses
(-) Difficult interpretable
(+) Easily trained

## Proposed




(+) Only escitory synapses
(+) No-canceling synapses
(+) Easily interpretable
(-) Constraints should be applied on training


Typical Representation
Input
Weights 1
Extracted Features
Weights 2
Prediction
Part-based Representation
Input
Weights 1
Extracted Features
Weights 2
Prediction

# EXPERIMENTAL RESULTS

- We experimentally evaluate the proposed method on Cartpole.
- Both actor and critic applied to 10-neuron linear layers, employing ReLU2 in the hidden layer.
- Each episode runs for 195 steps.
- We report the average accumulated reward and action probabilities of 5 training runs.

We compare the proposed method with two baselines using two different initialization schemes.

- Both schemes draw values from a Gaussian distribution $\theta \sim \mathcal{N}(0, \sigma_k)$ actor parameters given a distribution
- **Xavier/Glorot initialization** scheme:
  - $$\sigma_{xavier} = \sqrt{\frac{2}{n+m}}$$
- **He/Kaiming Initialization** scheme:
  - $$\sigma_{he} = \sqrt{2}\sqrt{\frac{2}{n+m}}$$

  Where n and m are the fan-in and fan-out of the layer, respectively.

The baselines optimizes the actor network app ling an existing in bibliography sign-preserving optimization method[3], named **Clipping Stochastic Gradient Ascent (CSGA)**.

$$\theta = \mathsf{max}\left(0, \theta_{\mathsf{old}} + \eta \frac{\partial \mathsf{L}^{\mathsf{actor}}}{\partial \theta_{\mathsf{old}}}\right).$$

[3]Chorowski, Jan, and Jacek M. Zurada. "Learning understandable neural networks with nonnegative weight constraints." IEEE transactions on neural networks and learning systems 26.1 (2014): 62-69.

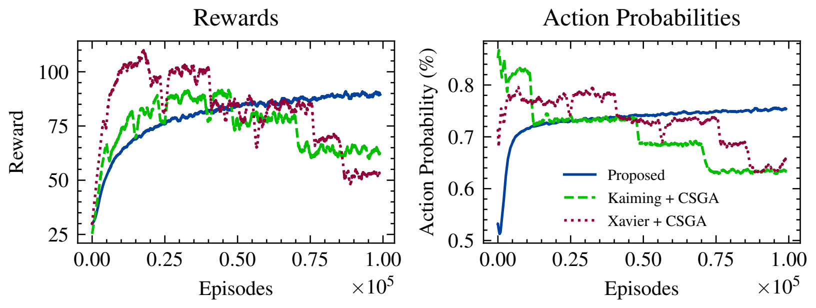



**Figure:** On the left, the figure depicts the obtained reward during training that is smoothed using a moving average filter with a window of 100. On the right, the action probabilities for each method are depicted using the same moving average setting.

Table: Average and variance of rewards both for training and evaluation phase over 5 runs.

| Method | Training | Evaluation |
|---|---|---|
| CSGA (Kaiming Init.) | $62.83 \pm 39.64$ | $89 \pm 98.59$ |
| CSGA (Xavier Init.) | $53.67 \pm 35.47$ | $58.2 \pm 78.4$ |
| **Proposed** | $\mathbf{89.45 \pm 1.04}$ | $\mathbf{140.4 \pm 43.9}$ |

**Baselines** evaluation indicates that:

- They are highly unstable.
- Resulting in poor local minimum.
- End up in significantly lower results.

**Baselines** evaluation indicates that:

- They are highly unstable.
- Resulting in poor local minimum.
- End up in significantly lower results.

**The proposed method** sufficiently demonstrates that:

- Builds robust model, resulting in consistent training.
- Achieving significantly higher performance than the baselines.

**Baselines optimization:**

$$\theta = \max\left(0, \theta_{\text{old}} + \eta \frac{\partial \mathrm{L}^{\text{actor}}}{\partial \theta_{\text{old}}}\right).$$

- Clipping method zeros out synapses when they try to change sign.
- Reducing the learning capacity of the model.
- Lead to vanishing gradient phenomena.
- Results in bad local minima or even halt the training process

**Baselines optimization:**

$$\theta = \max\left(0, \theta_{\text{old}} + \eta \frac{\partial L^{\text{actor}}}{\partial \theta_{\text{old}}}\right).$$

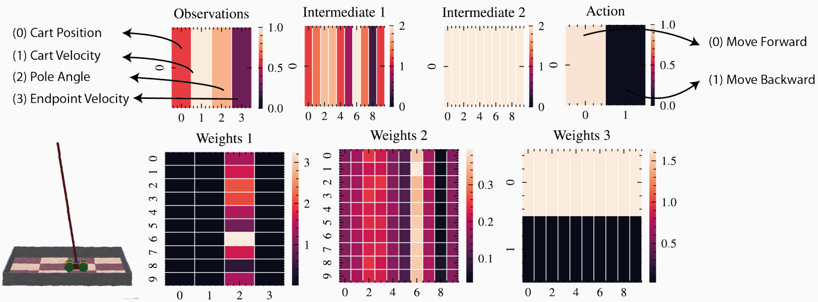
- Clipping method zeros out synapses when they try to change sign.
- Reducing the learning capacity of the model.
- Lead to vanishing gradient phenomena.
- Results in bad local minima or even halt the training process

**The proposed optimization:**

$$\theta = \left|\theta_{\text{old}} + \eta \frac{\partial L^{\text{actor}}}{\partial \theta_{\text{old}}}\right|$$

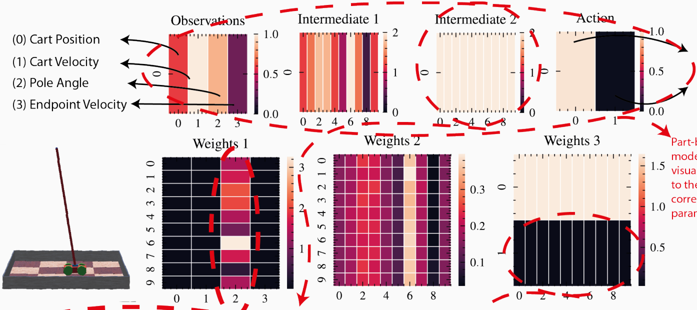
- Parameters remain non-negative without suppressing weights to zero.
- Allowing gradients to flow through the network since the absolute value operator has a non-zero derivative.
- Provides a smooth training process and consistent results

Pole is falling the front side of the cart



# FORWARD INTERPRETATION

Pole is falling the front side of the cart



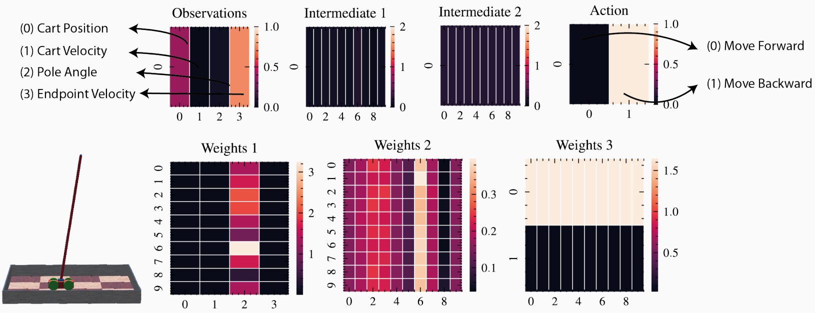

Part-based representation models defines an clear visual path from observation to the actions and visual correlations between inputs, parameters and outputs.

Weights values for inputs 0,1 and 3 are zeros, denoting that the most significant contribution comes from the pole angle

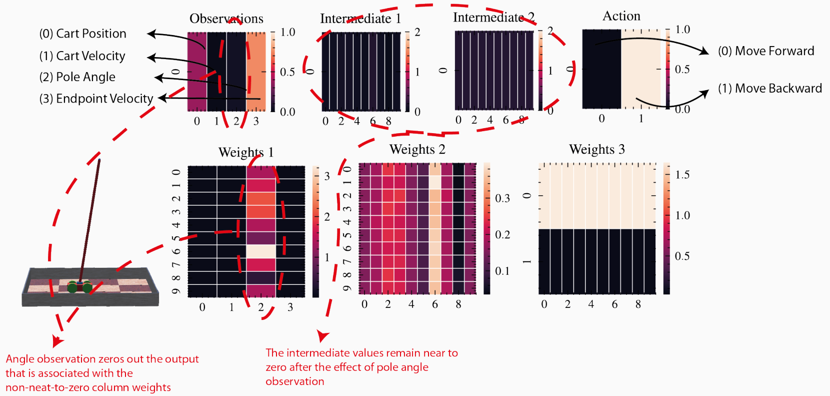
High values inticates the confidence of the action

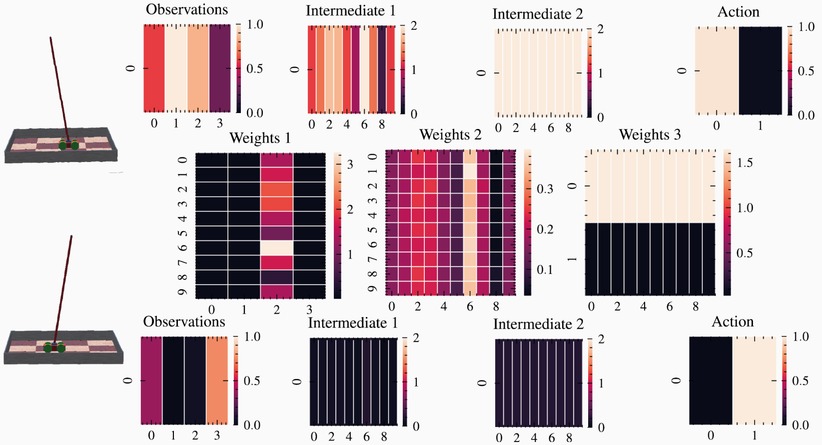
Zeros out the move backward output

Pole is falling the rear side of the cart

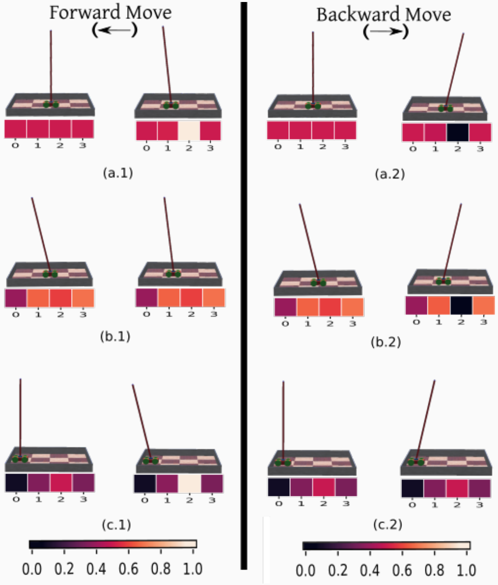


Pole is falling the rear side of the cart




Observations
Intermediate 1
Intermediate 2
Action
Weights 1
Weights 2
Weights 3
Observations
Intermediate 1
Intermediate 2
Action

# BACKWARD INTERPRETATION



# CONCLUSION & FUTURE WORKS

## CONCLUSIONS

1. The proposed approach **enables the extraction of part-based representations**.
2. Part-based representation **enhanced interpretability**.
3. To achieve this objective, the proposed method employs a non-negative initialization technique, followed by a modified sign-preserving training method.
4. **Enhancing training stability**.

**The proposed pipeline enables more efficient training of inherently explainable models based on the non-negative part-based representation of the actor.**

The promising results reported in this paper highlight several interesting future research directions.

- The proposed method **can also be extended to handle value-based RL approaches**, such as DQN.
- **Part-based representation learning to the critic model** could also provide further insight into the training dynamics of the RL process, potentially leading to more robust algorithms.
- **Combining the proposed method with distillation approaches**, could potentially allow for better guidance of the optimization process and learning more accurate policies.

This work was supported by the European Union's Horizon 2020 Research and Innovation Program (**OpenDR**) under Grant 871449. This publication reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information it contains.

Project Site: `https://opendr.eu`

Thank you!

Questions?