Playing Atari games with an Interpretable Agent

Erwan Lecarpentier, Dennis G. Wilson, Sylvain Cussat-Blanc, Hervé Luga, Guillaume Jubelin, and, Lionel Cordesses

November 9, 2021

Content

Content

1. [Context] Interpretability in Atari

Content

1. [Context] Interpretability in Atari
2. [Method] Cartesian Genetic Programming

Content

1. [Context] Interpretability in Atari
2. [Method] Cartesian Genetic Programming
3. [Experiments] Results on Atari

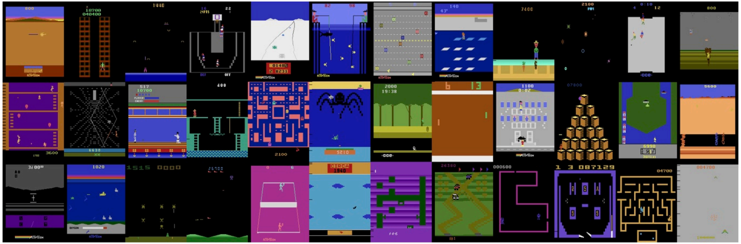
Content

1. [Context] Interpretability in Atari
2. [Method] Cartesian Genetic Programming
3. [Experiments] Results on Atari
4. [Conclusion] Next steps

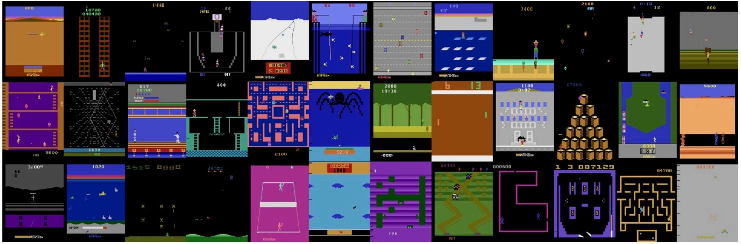
Content

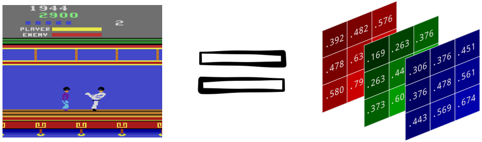
[Context] Playing Atari games (from pixels input)

[Context] Playing Atari games (from pixels input)

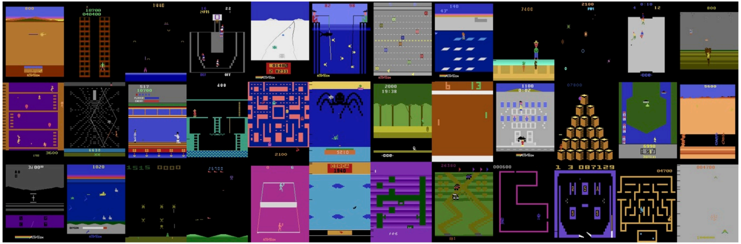


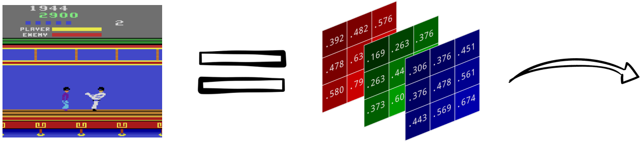
[Context] Playing Atari games (from pixels input)





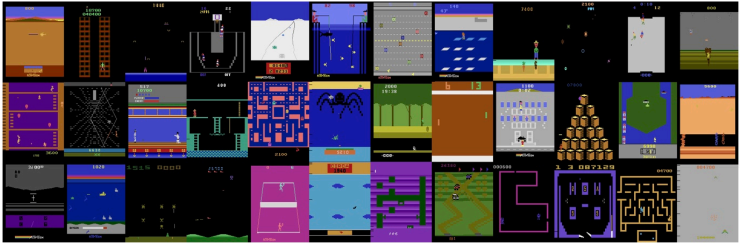
[Context] Playing Atari games (from pixels input)





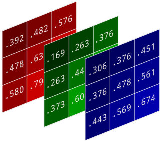
$a \in \mathcal{A}$

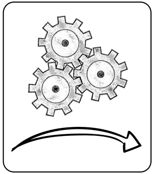
[Context] Playing Atari games (from pixels input)



Agent

$=$





$a \in \mathcal{A}$

[Context] Interpretability

Interpretability is the degree to which a human can understand the cause of a decision[1].

[1] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence 267 (2019): 1-38

[Context] Interpretability in Machine Learning

[Context] Interpretability in Machine Learning

Interpretability

"Techniques for Interpretable Machine Learning", Du, Liu, and Hu 2019

[Context] Interpretability in Machine Learning

Interpretability	**Inherent** *Self-explained model ...*	**Post-hoc** *Explain model with another model ...*

"Techniques for Interpretable Machine Learning", Du, Liu, and Hu 2019

[Context] Interpretability in Machine Learning

Interpretability	**Inherent** *Self-explained model ...*	**Post-hoc** *Explain model with another model ...*
Global *... for ALL decisions*		
Local *... for SOME decisions*		

"Techniques for Interpretable Machine Learning", Du, Liu, and Hu 2019

[Context] Interpretability in Machine Learning

Interpretability	**Inherent** *Self-explained model …*	**Post-hoc** *Explain model with another model …*
Global *… for ALL decisions*	Linear model	
Local *… for SOME decisions*		

"Techniques for Interpretable Machine Learning", Du, Liu, and Hu 2019

[Context] Interpretability in Machine Learning

Interpretability	**Inherent** *Self-explained model ...*	**Post-hoc** *Explain model with another model ...*
Global *... for ALL decisions*	Linear model	Learn activating input images in CNN
Local *... for SOME decisions*		

"Techniques for Interpretable Machine Learning", Du, Liu, and Hu 2019

[Context] Interpretability in Machine Learning

Interpretability	**Inherent** *Self-explained model ...*	**Post-hoc** *Explain model with another model ...*
Global *... for ALL decisions*	Linear model	Learn activating input images in CNN
Local *... for SOME decisions*	Attention mechanisms	

"Techniques for Interpretable Machine Learning", Du, Liu, and Hu 2019

[Context] Interpretability in Machine Learning

Interpretability	**Inherent** *Self-explained model ...*	**Post-hoc** *Explain model with another model ...*
Global *... for ALL decisions*	Linear model	Learn activating input images in CNN
Local *... for SOME decisions*	Attention mechanisms	Local approximation with white-box model

"Techniques for Interpretable Machine Learning", Du, Liu, and Hu 2019

1. Interpretability in Atari

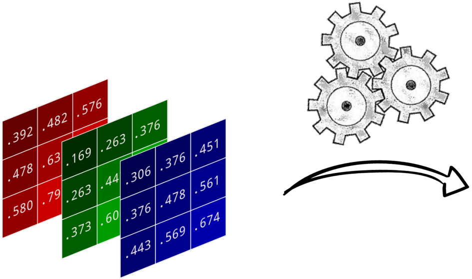
1. Interpretability in Atari

Why did you take the action “kick”?

1. Interpretability in Atari

Why did you take the action "kick"?

Because:



$$a = \text{kick}$$

1. Interpretability in Atari

Why did you take the action “kick”?

Because:

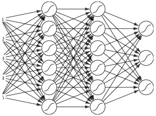
$\hat{M}\Big(a = \text{kick} \;\Big|$  $\Big) \geq \hat{M}\Big(a \neq \text{kick} \;\Big|$  $\Big)$

1. Interpretability in Atari

Why did you take the action "kick"?

Because:

$\hat{M}\Big(a = \text{kick} \;\Big|\;$  $\Big) \geq \hat{M}\Big(a \neq \text{kick} \;\Big|\;$  $\Big)$



1. Interpretability in Atari

Why did you take the action “kick”?

Because:

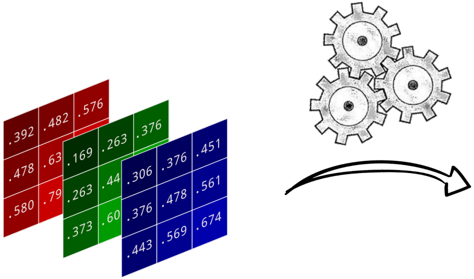
$$\sigma (0.403 \times 0.635 + 0.472 \times 0.687 + 0.281 \times 0.53 + 0.866 \times 0.931 + 0.182 \times 0.427 + 0.834 \times 0.913 + \sigma (0.986 \times 0.993 + 0.169 \times 0.412) + 0.755 \times 0.869 + 0.352 \times 0.593 + 0.366 \times 0.605) + \sigma (0.662 \times 0.813 + 0.639 \times 0.8 + 0.281 \times 0.53 + 0.516 \times 0.718 + 0.187 \times 0.432) + \sigma (0.867 \times 0.931 + 0.917 \times 0.958 + 0.793 \times 0.89 + 0.393 \times 0.627 + 0.281 \times 0.531 + 0.5 \times 0.707 + 0.772 \times 0.879) + \sigma (0.854 \times 0.924 + 0.411 \times 0.641 + 0.052 \times 0.228 + \sigma (0.712 \times 0.844 + 0.959 \times 0.979) + 0.197 \times 0.444 + 0.456 \times 0.675 + 0.785 \times 0.886) + \sigma (0.72 \times 0.849 + 0.998 \times 0.999 + 0.216 \times 0.465 + 0.034 \times 0.184 + 0.003 \times 0.058 + 0.55 \times 0.741 + 0.949 \times 0.974 + 0.815 \times 0.903) + \sigma (0.768 \times 0.876 + 0.494 \times 0.703 + 0.838 \times 0.915) + \sigma (0.153 \times 0.391 + 0.103 \times 0.322 + 0.344 \times 0.587 + 0.136 \times 0.369 + 0.115 \times 0.339 + 0.295 \times 0.543 + 0.656 \times 0.81 + 0.04 \times 0.2) + \sigma (0.403 \times 0.635 + 0.472 \times 0.687 + 0.281 \times 0.53 + 0.866 \times 0.931 + 0.182 \times 0.427 + 0.834 \times 0.913 + \sigma (0.986 \times 0.993 + 0.169 \times 0.412) + 0.755 \times 0.869 + 0.352 \times 0.593 + 0.366 \times 0.605) + \sigma (0.662 \times 0.813 + 0.639 \times 0.8 + 0.281 \times 0.53 + 0.516 \times 0.718 + 0.187 \times 0.432) + \sigma (0.867 \times 0.931 + 0.917 \times 0.958 + 0.793 \times 0.89 + 0.393 \times 0.627 + 0.281 \times 0.531 + 0.5 \times 0.707 + 0.772 \times 0.879) + \sigma (0.854 \times 0.924 + 0.411 \times 0.641 + 0.052 \times 0.228 + \sigma (0.712 \times 0.844 + 0.959 \times 0.979) + 0.197 \times 0.444 + 0.456 \times 0.675 + 0.785 \times 0.886) + \sigma (0.72 \times 0.849 + 0.998 \times 0.999 + 0.216 \times 0.465 + 0.034 \times 0.184 + 0.003 \times 0.058 + 0.55 \times 0.741 + 0.949 \times 0.974 + 0.815 \times 0.903) + \sigma (0.768 \times 0.876 + 0.494 \times 0.703 + 0.838 \times 0.915) + \sigma (0.153 \times 0.391 + 0.103 \times 0.322 +$$

$$\geq$$

$$\sigma (0.662 \times 0.813 + 0.639 \times 0.8 + 0.281 \times 0.53 + 0.516 \times 0.718 + 0.187 \times 0.432) + \sigma (0.867 \times 0.931 + 0.917 \times 0.958 + 0.793 \times 0.89 + 0.393 \times 0.627 + 0.281 \times 0.531 + 0.5 \times 0.707 + 0.772 \times 0.879) + \sigma (0.854 \times 0.924 + 0.411 \times 0.641 + 0.052 \times 0.228 + \sigma (0.712 \times 0.844 + 0.959 \times 0.979) + 0.197 \times 0.444 + 0.456 \times 0.675 + 0.785 \times 0.886) + \sigma (0.72 \times 0.849 + 0.998 \times 0.999 + 0.216 \times 0.465 + 0.034 \times 0.184 + 0.003 \times 0.058 + 0.55 \times 0.741 + 0.949 \times 0.974 + 0.815 \times 0.903) + \sigma (0.768 \times 0.876 + 0.494 \times 0.703 + 0.838 \times 0.915) + \sigma (0.403 \times 0.635 + 0.472 \times 0.687 + 0.281 \times 0.53 + 0.866 \times 0.931 + 0.182 \times 0.427 + 0.834 \times 0.913 + \sigma (0.986 \times 0.993 + 0.169 \times 0.412) + 0.755 \times 0.869 + 0.352 \times 0.593 + 0.366 \times 0.605) + \sigma (0.153 \times 0.391 + 0.103 \times 0.322 + 0.344 \times 0.587 + 0.136 \times 0.369 + 0.115 \times 0.339 + 0.295 \times 0.543 + 0.656 \times 0.81 + 0.04 \times 0.2) + \sigma (0.867 \times 0.931 + 0.917 \times 0.958 + 0.793 \times 0.89 + 0.393 \times 0.627 + 0.281 \times 0.531 + 0.5 \times 0.707 + 0.772 \times 0.879) + \sigma (0.854 \times 0.924 + 0.411 \times 0.641 + 0.052 \times 0.228 + \sigma (0.712 \times 0.844 + 0.959 \times 0.979) + 0.197 \times 0.444 + 0.456 \times 0.675 + 0.785 \times 0.886) + \sigma (0.72 \times 0.849 + 0.998 \times 0.999 + 0.216 \times 0.465 + 0.034 \times 0.184 + 0.003 \times 0.058 + 0.55 \times 0.741 + 0.949 \times 0.974 + 0.815 \times 0.903) + \sigma (0.768 \times 0.876 + 0.494 \times 0.703 + 0.838 \times 0.915) + \sigma (0.403 \times 0.635 + 0.472 \times 0.687 + 0.281 \times 0.53 + 0.866 \times 0.931 + 0.182 \times 0.427 + 0.834 \times 0.913 + \sigma (0.986 \times 0.993 + 0.169 \times 0.412) + 0.755 \times 0.869 + 0.352 \times 0.593 + 0.366 \times 0.605) + \sigma (0.153 \times 0.391 + 0.103 \times 0.322 + 0.344 \times 0.587 + 0.136 \times 0.369 + 0.115 \times 0.339 + 0.295 \times 0.543 + 0.656 \times 0.81 + 0.04 \times 0.2)$$

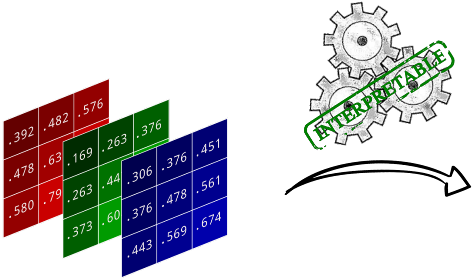
[Context] Goal: Interpretable Agent

[Context] Goal: Interpretable Agent



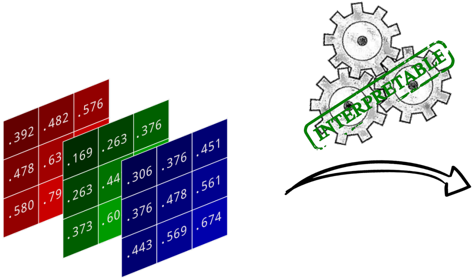
$$a \in \mathcal{A}$$

[Context] Goal: Interpretable Agent

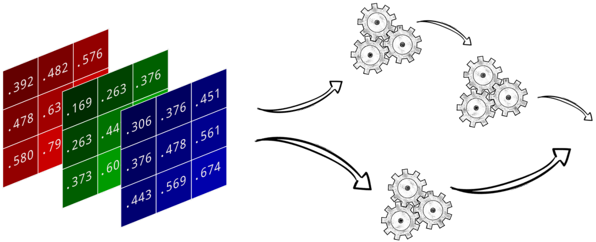


$$a \in \mathcal{A}$$

[Context] Goal: Interpretable Agent

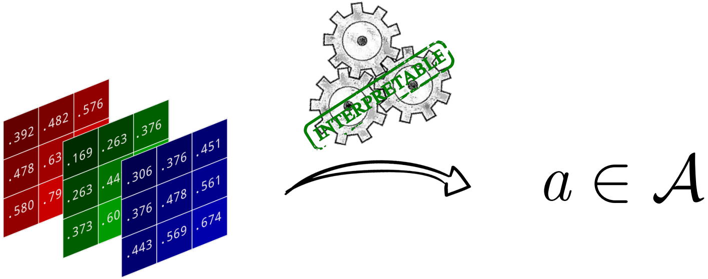


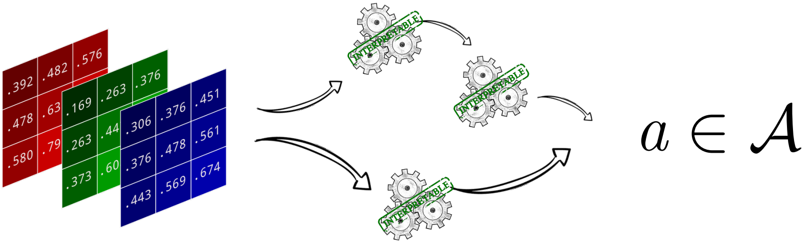
$$a \in \mathcal{A}$$



$$a \in \mathcal{A}$$

[Context] Goal: Interpretable Agent





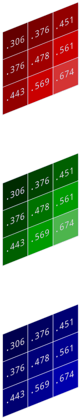
Content

1. [Context] Interpretability in Atari
2. [Method] Cartesian Genetic Programming
3. [Experiments] Results on Atari
4. [Conclusion] Next steps

[Method] Approach: Interpretable Encoder – Controller

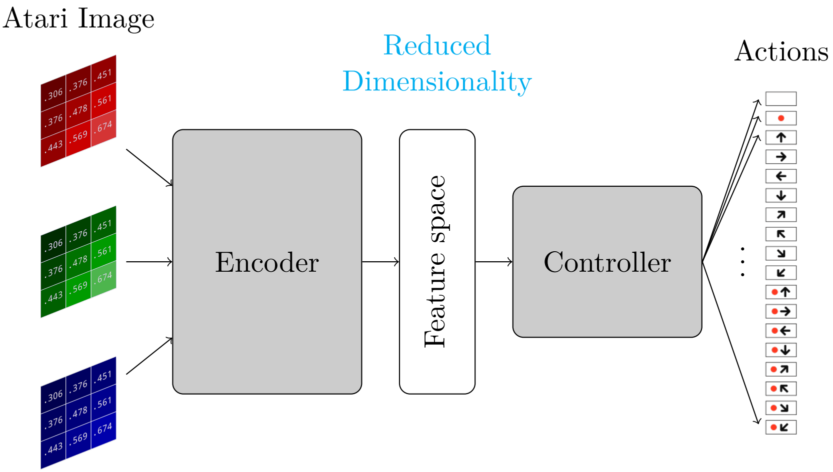
[Method] Approach: Interpretable Encoder – Controller

Atari Image

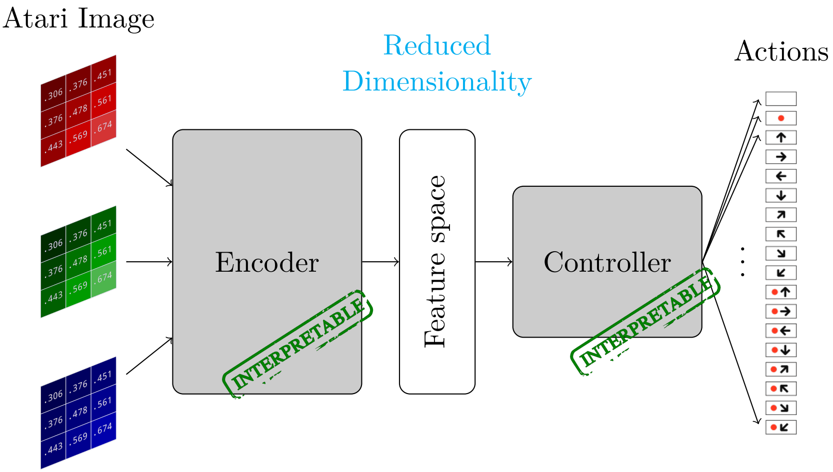


Actions

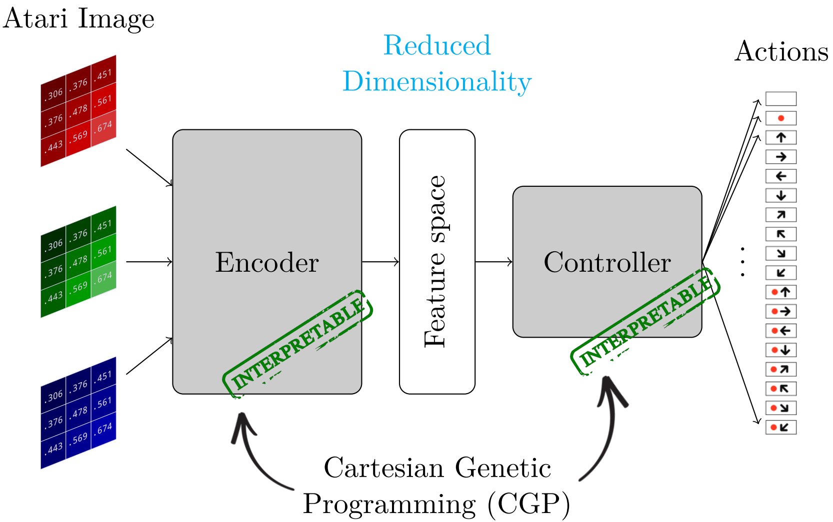
[Method] Approach: Interpretable Encoder – Controller



[Method] Approach: Interpretable Encoder – Controller

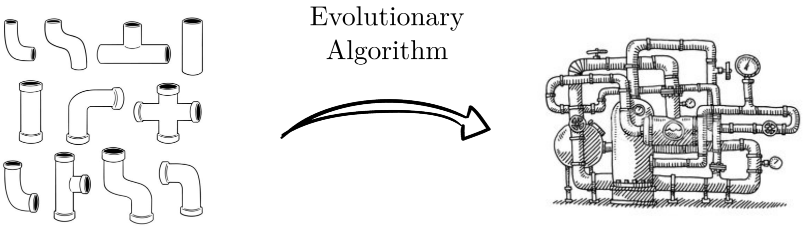


[Method] Approach: Interpretable Encoder – Controller

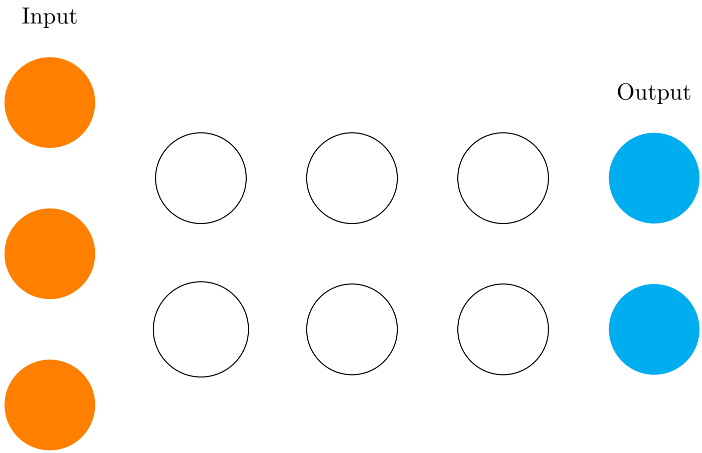


[Method] Cartesian Genetic Programming (CGP)

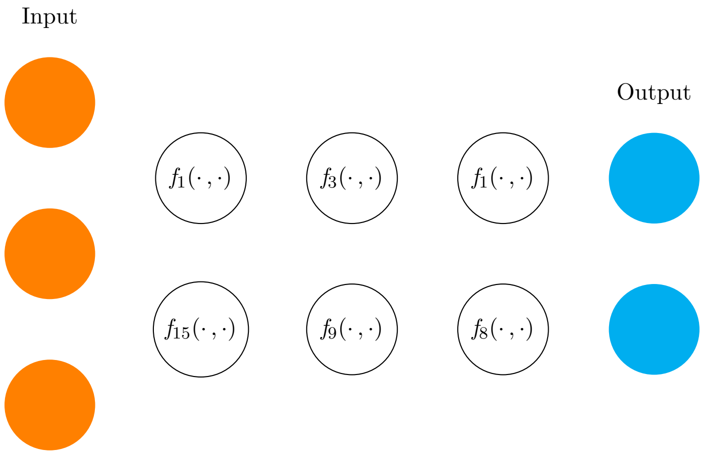
[Method] Cartesian Genetic Programming (CGP)



[Method] Cartesian Genetic Programming (CGP)

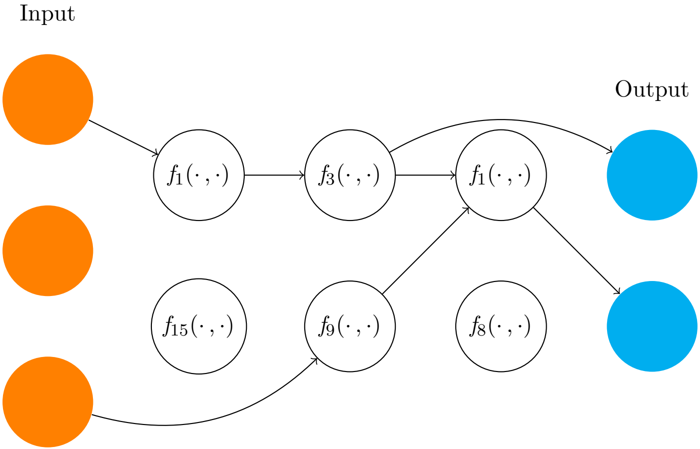


[Method] Cartesian Genetic Programming (CGP)



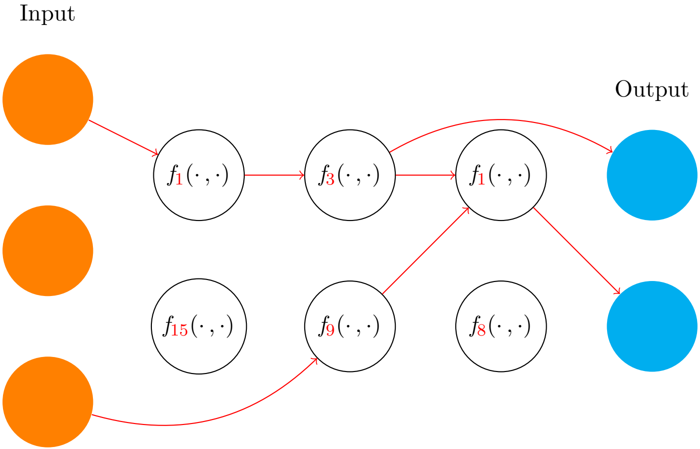
Function pool: $\{f_i : \mathcal{X}^2 \to \mathcal{X}\}_i$

[Method] Cartesian Genetic Programming (CGP)



Function pool: $\{f_i : \mathcal{X}^2 \to \mathcal{X}\}_i$

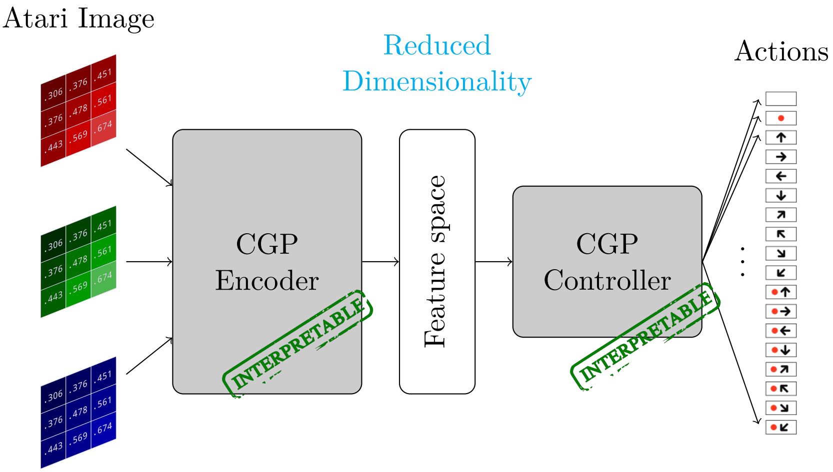
[Method] Cartesian Genetic Programming (CGP)



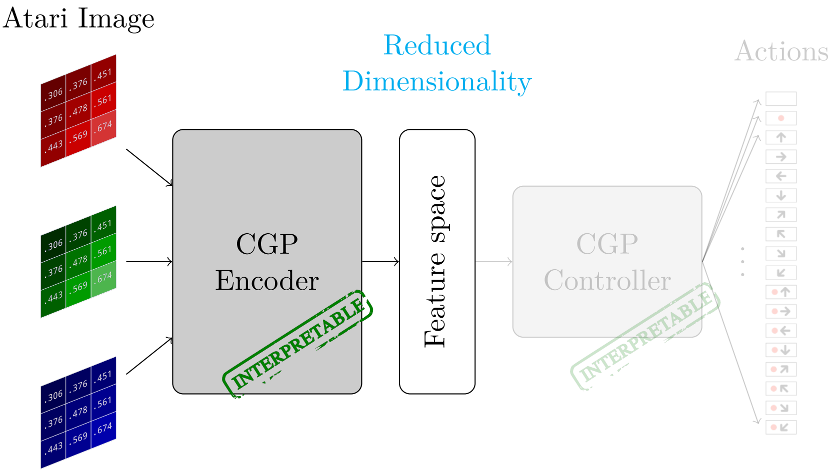
Function pool: $\{f_i : \mathcal{X}^2 \to \mathcal{X}\}_i$

Genotype: $[1, 1, 2, 13, 1, \ldots, 3] \in \mathbb{N}^{3\times\text{number of nodes} + \text{number of outputs}}$

[Method] Approach: Interpretable Encoder – Controller

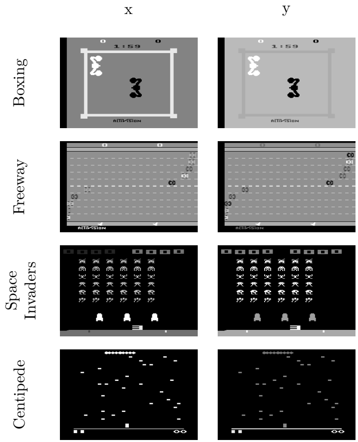


[Method] Approach: Interpretable Encoder – Controller

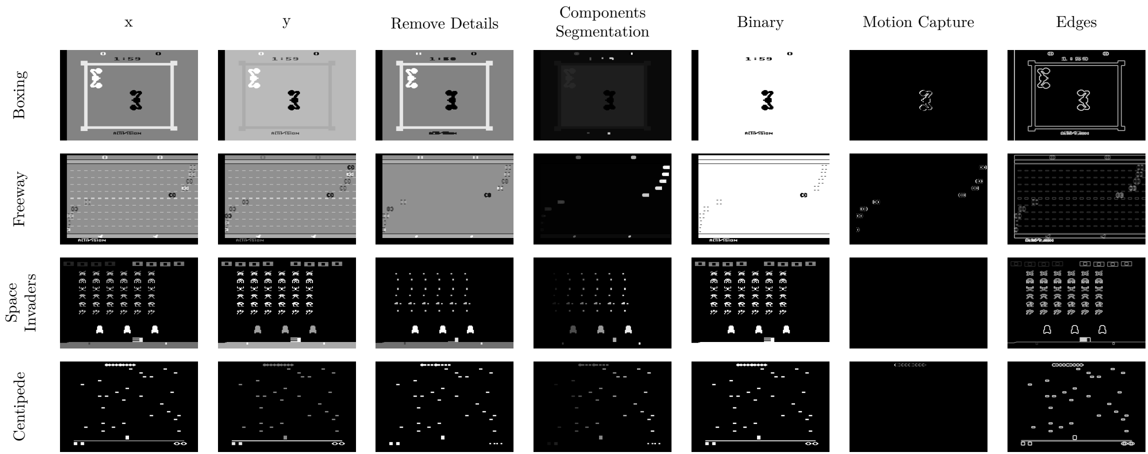


[Method] Encoder's Functions

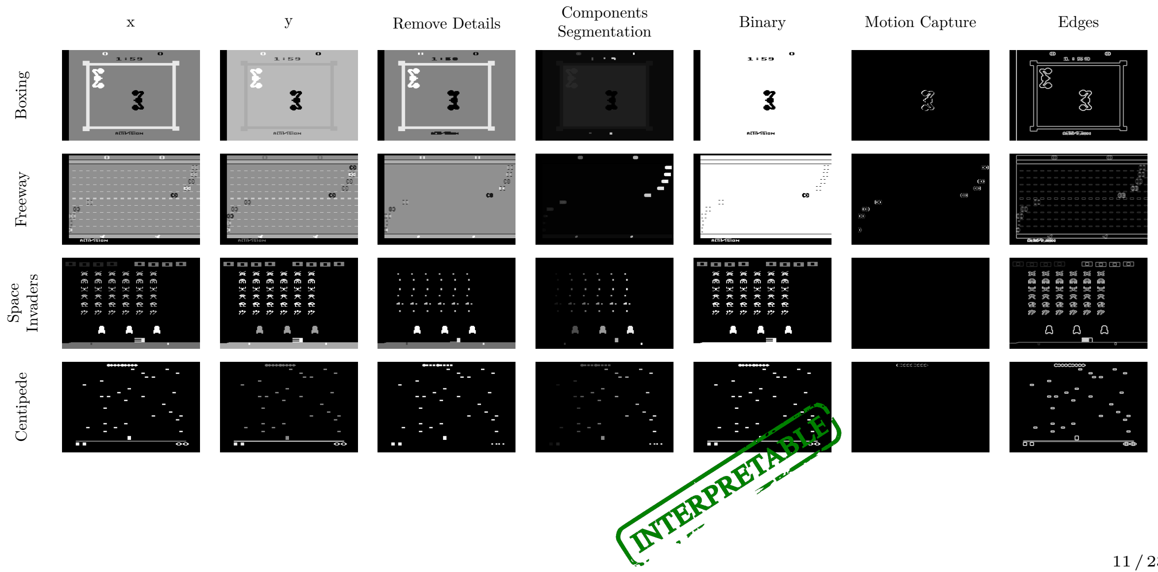
[Method] Encoder’s Functions



[Method] Encoder's Functions



[Method] Encoder's Functions

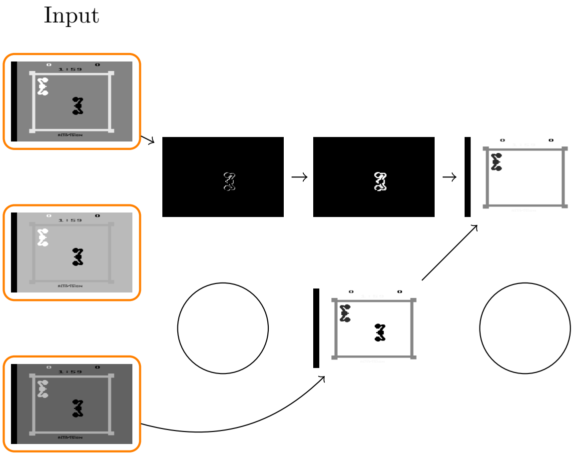


[Method] Encoder

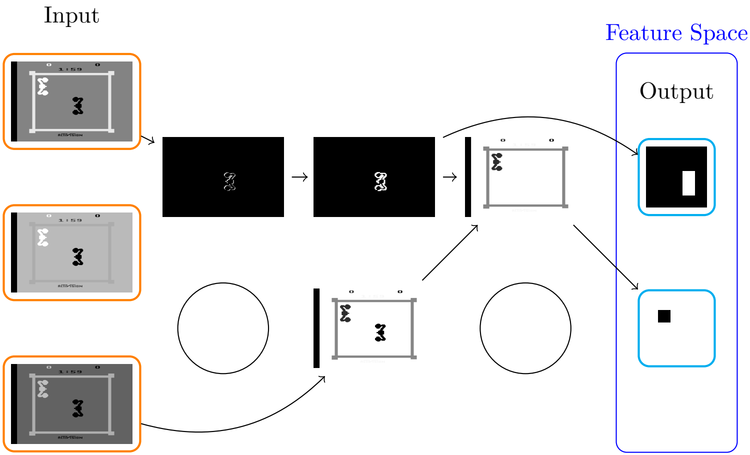
[Method] Encoder

Input

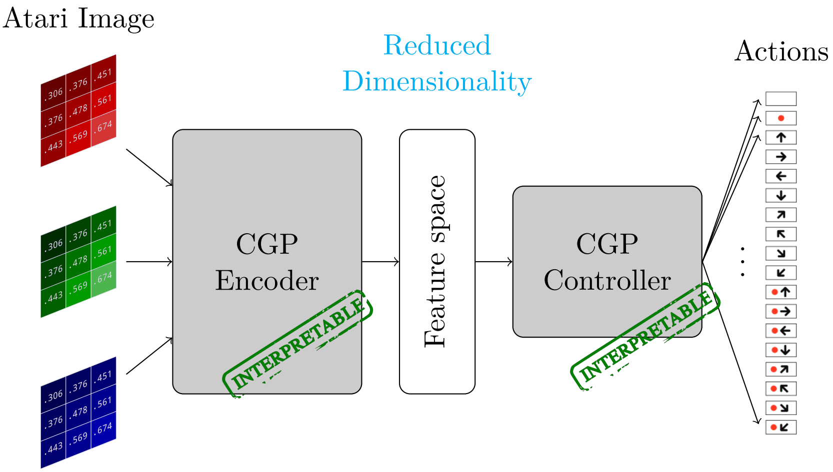
[Method] Encoder



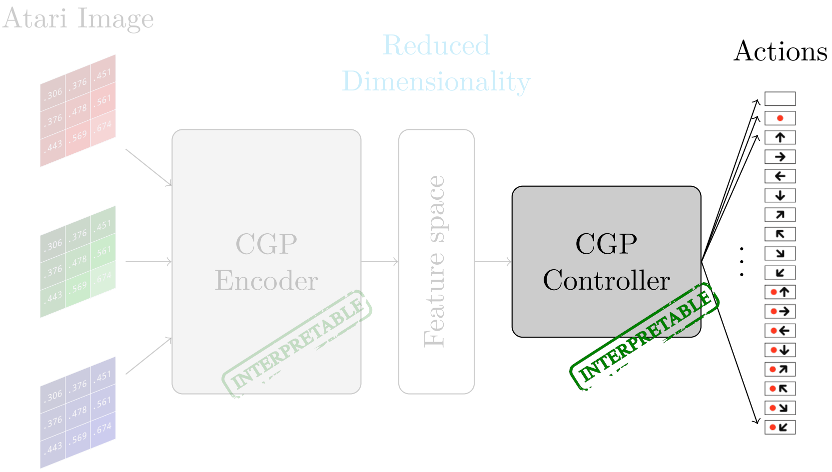
[Method] Encoder



[Method] Approach: Interpretable Encoder – Controller



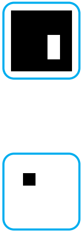
[Method] Approach: Interpretable Encoder – Controller



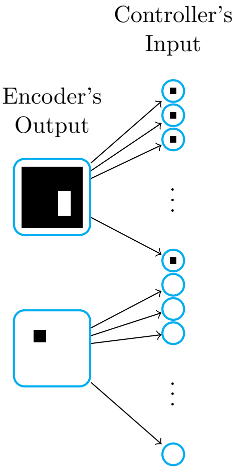
[Method] Controller

[Method] Controller

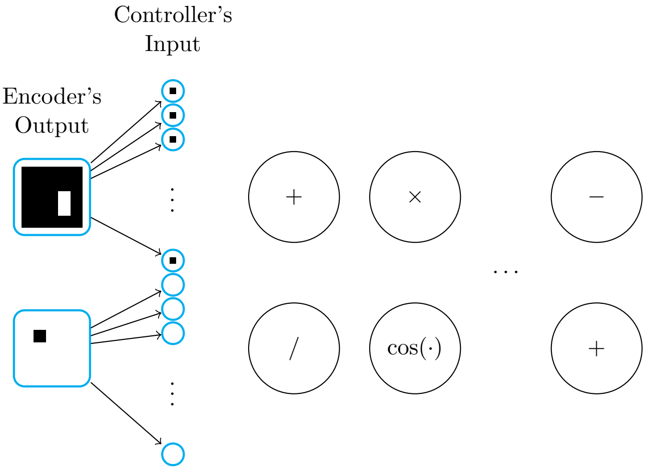
Encoder's
Output



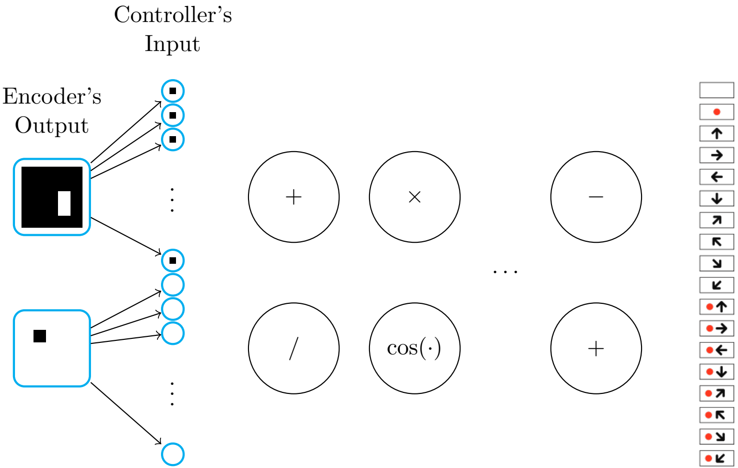
[Method] Controller



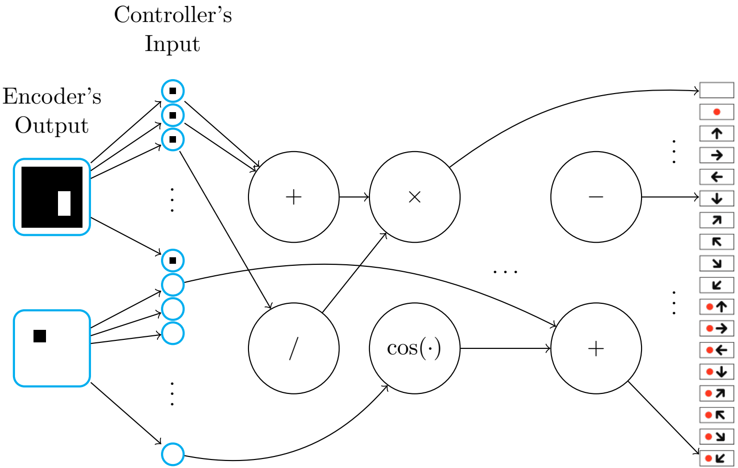
[Method] Controller



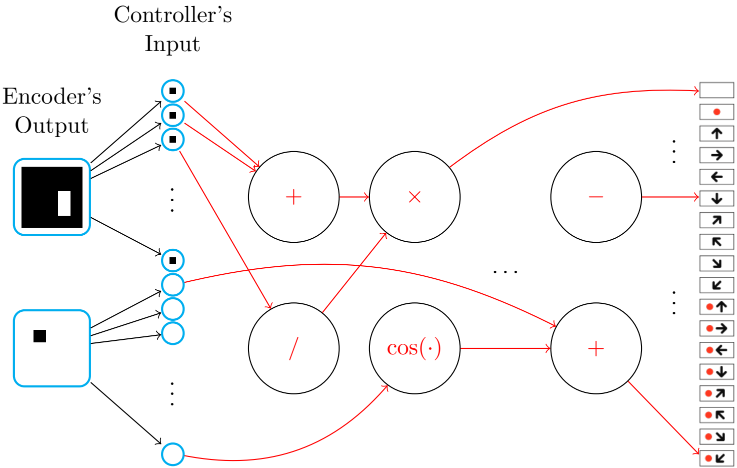
[Method] Controller



[Method] Controller



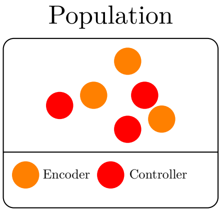
[Method] Controller



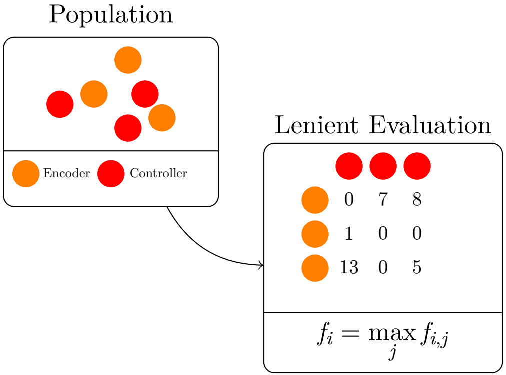
Genotype: $[1, 1, 2, 13, 1, \ldots, 3] \in \mathbb{N}^{3 \times \text{number of nodes} + \text{number of outputs}}$

[Method] Optimizer: $1 + \lambda$ co-evolution

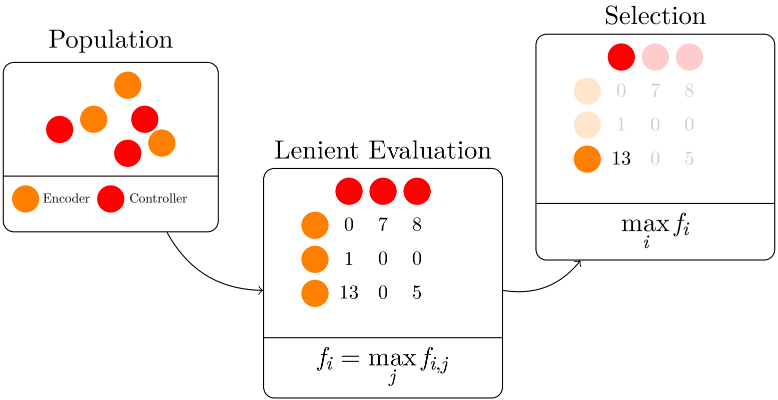
[Method] Optimizer: $1 + \lambda$ co-evolution



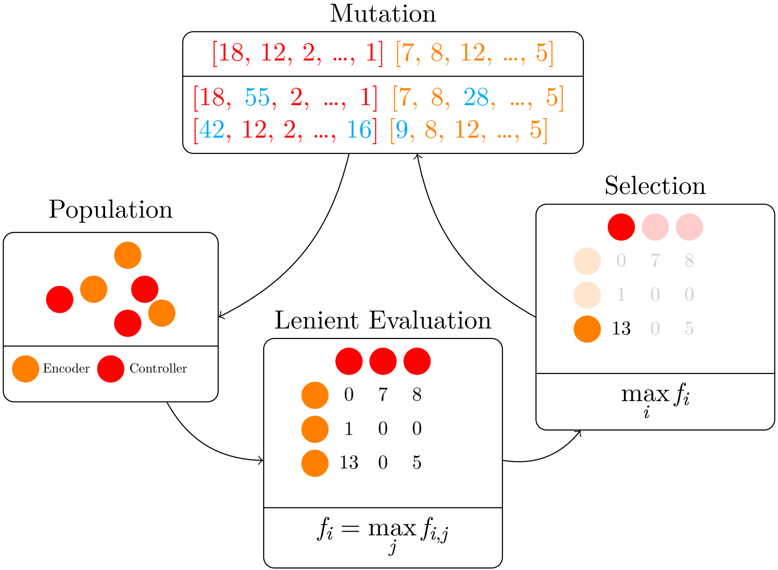
[Method] Optimizer: $1 + \lambda$ co-evolution



[Method] Optimizer: $1 + \lambda$ co-evolution



[Method] Optimizer: $1 + \lambda$ co-evolution



Content

1. [Context] Interpretability in Atari
2. [Method] Cartesian Genetic Programming
3. [Experiments] Results on Atari
4. [Conclusion] Next steps

[Experiments] Settings

[Experiments] Settings

- **Encoder input:** 1 gray-level down-scaled image

[Experiments] Settings

- **Encoder input:** 1 gray-level down-scaled image
- **Stochasticity:** `repeat_action_probability = 0.25`

[Experiments] Settings

- **Encoder input:** 1 gray-level down-scaled image
- **Stochasticity:** `repeat_action_probability = 0.25`
- **Fitness evaluation:** score obtained after 1 roll-out

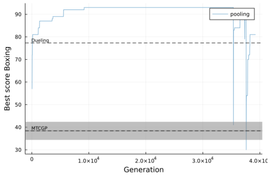
[Experiments] Settings

- **Encoder input:** 1 gray-level down-scaled image
- **Stochasticity:** `repeat_action_probability = 0.25`
- **Fitness evaluation:** score obtained after 1 roll-out
- **Seed:** same seed for all evaluations

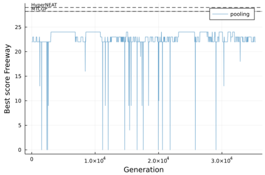
[Experiments] Results: performance

[Experiments] Results: performance

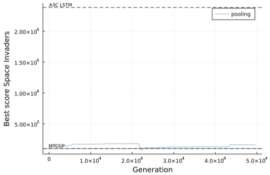
Boxing



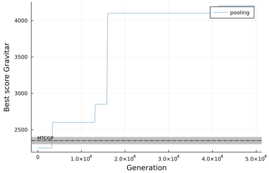
Freeway



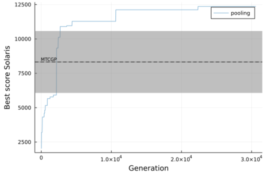
Space Invaders



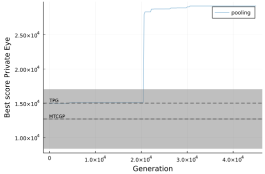
Gravitar



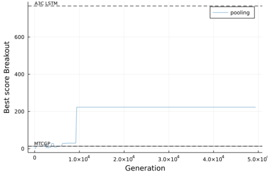
Solaris



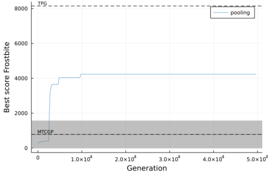
Private Eye



Breakout



Frostbite



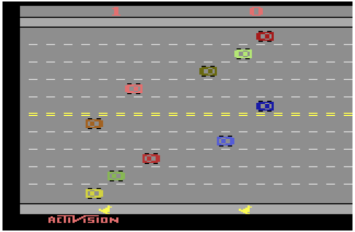
https://github.com/erwanlecarpentier/ICGP-results

[Experiments] Results: interpretability

[Experiments] Results: interpretability

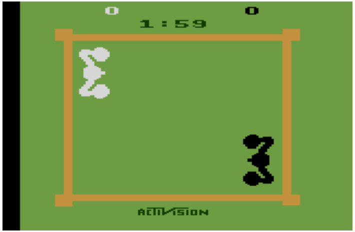
Videos:

`Freeway`



`Space Invaders`

`Boxing`



[Experiments] Stochastic setting

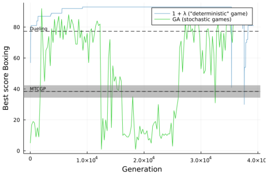
[Experiments] Stochastic setting

Same setting with a different seed for each generation

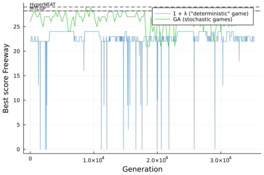
[Experiments] Stochastic setting

Same setting with a different seed for each generation

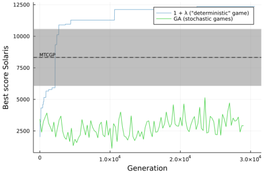
Boxing



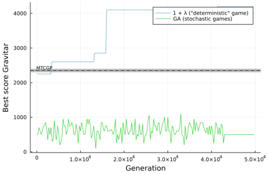
Freeway



Solaris



Gravitar

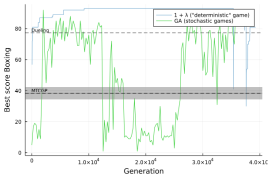


https://github.com/erwanlecarpentier/ICGP-results

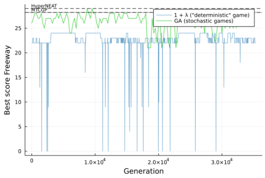
[Experiments] Stochastic setting

Same setting with a different seed for each generation

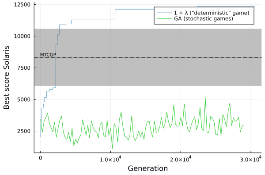
Boxing



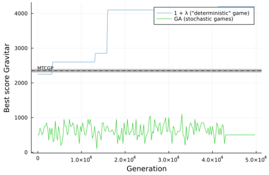
Freeway



Solaris



Gravitar



- Boxing: collapse
- Freeway Solaris Gravitar: no learning progress

https://github.com/erwanlecarpentier/ICGP-results

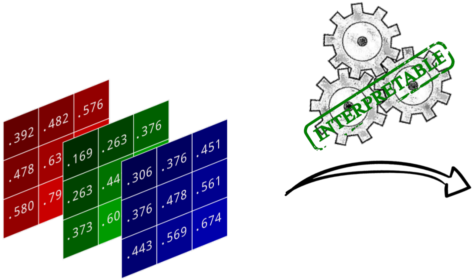
Content

1. [Context] Interpretability in Atari
2. [Method] Cartesian Genetic Programming
3. [Experiments] Results on Atari
4. [Conclusion] Next steps

[Conclusion]

[Conclusion]

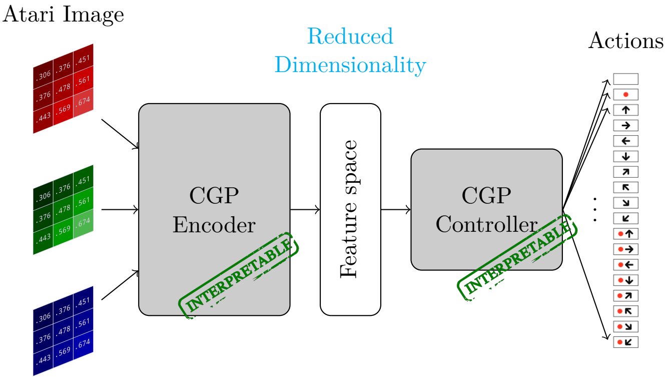
- Goal: interpretable agent in pixel-based Atari games



$a \in \mathcal{A}$

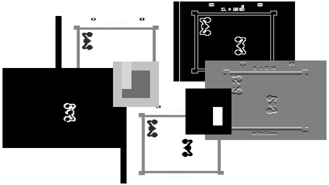
[Conclusion]

- Goal: interpretable agent in pixel-based Atari games
- Method: CGP co-evolution in an encoder-controller scheme



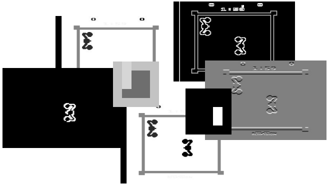
[Conclusion]

- Goal: interpretable agent in pixel-based Atari games
- Method: CGP co-evolution in an encoder-controller scheme
- Encoder: interpretable image processing functions



[Conclusion]

- Goal: interpretable agent in pixel-based Atari games
- Method: CGP co-evolution in an encoder-controller scheme
- Encoder: interpretable image processing functions
- Controller: interpretable scalar functions



cos

$\times \quad - \quad \div$

$+$

[Conclusion]

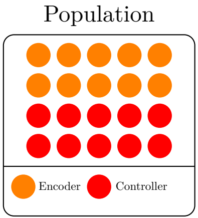
- Goal: interpretable agent in pixel-based Atari games
- Method: CGP co-evolution in an encoder-controller scheme
- Encoder: interpretable image processing functions
- Controller: interpretable scalar functions
- Experiments:

	Performance	Interpretability
Atari deterministic	OK	OK
Atari stochastic	NOT YET	OK

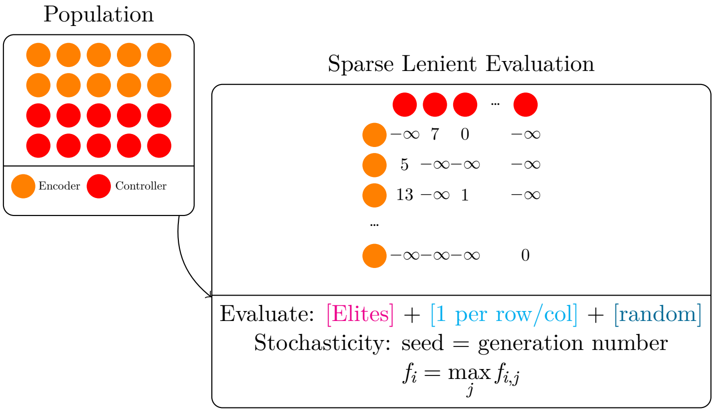
Images: pixabay.com and Wilson, Dennis G., et al. ”Evolving simple programs for playing Atari games.” GECCO 2018

GA co-evolution

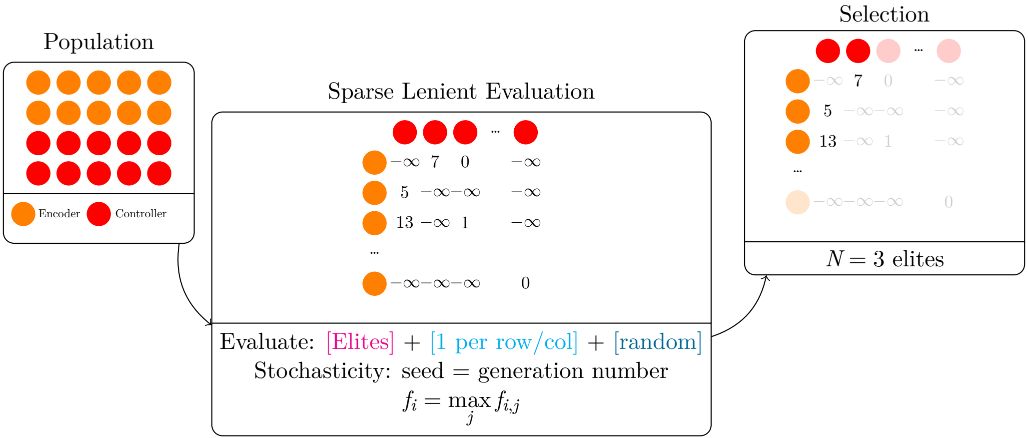
GA co-evolution



GA co-evolution



GA co-evolution



GA co-evolution