

Playing Atari games with an Interpretable Agent

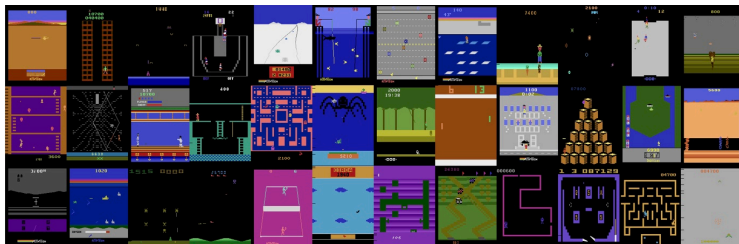
Erwan Lecarpentier, Dennis G. Wilson,
Sylvain Cussat-Blanc and Hervé Luga

June 3, 2021

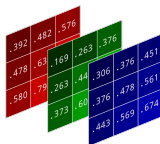
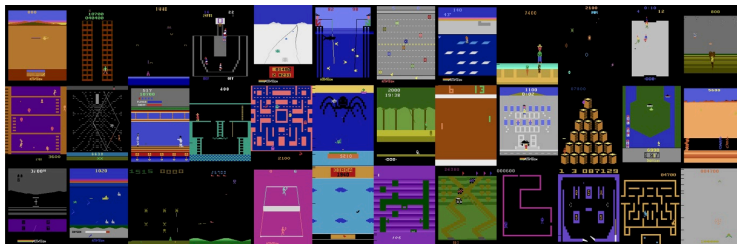


Playing Atari games (from pixels input)

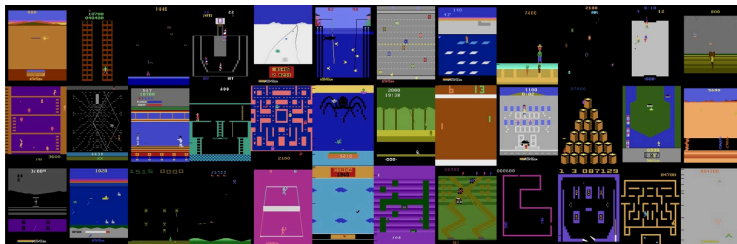
Playing Atari games (from pixels input)



Playing Atari games (from pixels input)

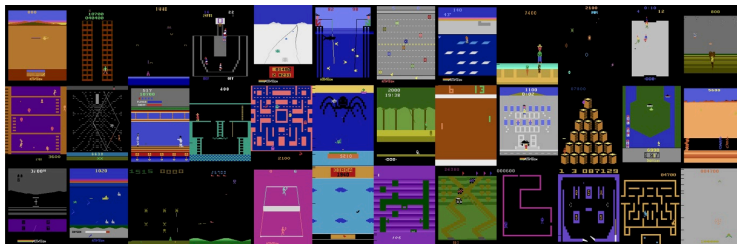


Playing Atari games (from pixels input)

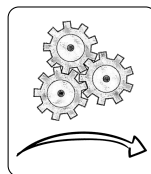


$a \in \mathcal{A}$

Playing Atari games (from pixels input)



Agent



$a \in \mathcal{A}$

Interpretability

Interpretability

Interpretability is the degree to which a human can understand the cause of a decision¹.

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

Interpretability in Atari

Interpretability in Atari



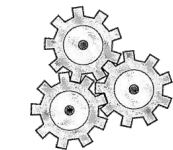
Why did you take
the action “kick”?

Interpretability in Atari



Why did you take
the action “kick”?

Because:



$a = \text{kick}$

Interpretability in Atari



Why did you take
the action “kick”?

Because:

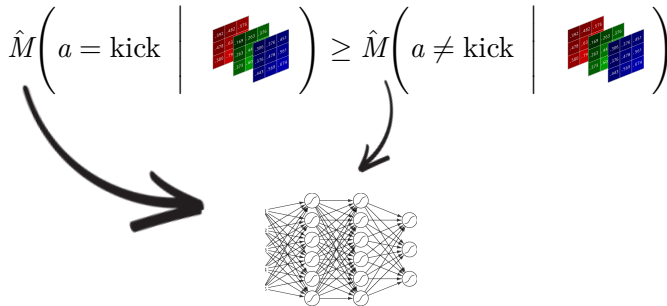
$$\hat{M}\left(a = \text{kick} \mid \begin{array}{|c|} \hline \text{Red} \\ \hline \end{array} \begin{array}{|c|} \hline \text{Green} \\ \hline \end{array} \begin{array}{|c|} \hline \text{Blue} \\ \hline \end{array}\right) \geq \hat{M}\left(a \neq \text{kick} \mid \begin{array}{|c|} \hline \text{Red} \\ \hline \end{array} \begin{array}{|c|} \hline \text{Green} \\ \hline \end{array} \begin{array}{|c|} \hline \text{Blue} \\ \hline \end{array}\right)$$

Interpretability in Atari



Why did you take
the action “kick”?

Because:



Interpretability in Atari



Why did you take the action “kick”?

Because:

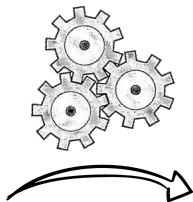
$$\begin{aligned} & \sigma(0.403 \times 0.635 + 0.472 \times 0.687 + 0.281 \times 0.53 + 0.866 \times 0.931 + 0.182 \\ & \times 0.427 + 0.834 \times 0.913 + \sigma(0.986 \times 0.993 + 0.169 \times 0.412) + 0.755 \times \\ & 0.869 + 0.352 \times 0.593 + 0.366 \times 0.605) + \sigma(0.662 \times 0.813 + 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 + \\ & 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) \end{aligned}$$

\geq

$$\begin{aligned} & \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) \end{aligned}$$

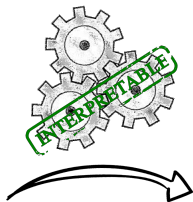
Goal: Interpretable Agent

Goal: Interpretable Agent



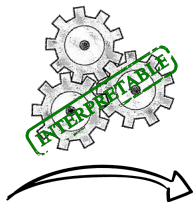
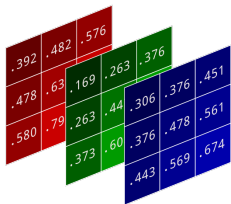
$$a \in \mathcal{A}$$

Goal: Interpretable Agent

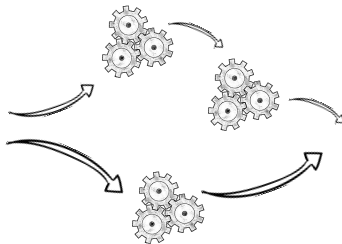
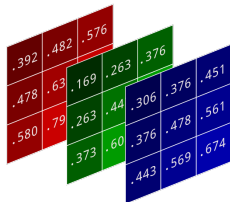


$$a \in \mathcal{A}$$

Goal: Interpretable Agent

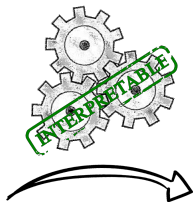
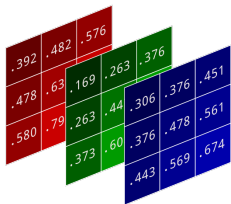


$$a \in \mathcal{A}$$

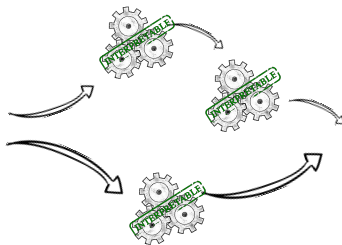
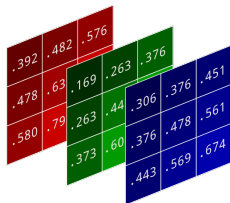


$$a \in \mathcal{A}$$

Goal: Interpretable Agent



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



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

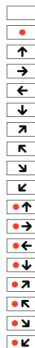
Approach: Interpretable Encoder – Controller

Approach: Interpretable Encoder – Controller

Atari Image

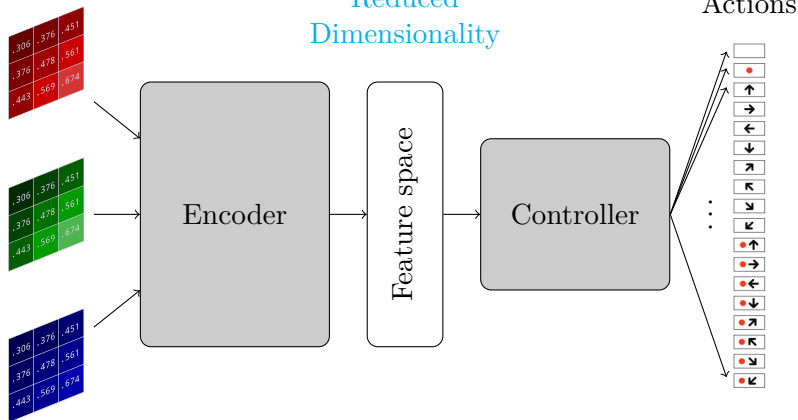


Actions



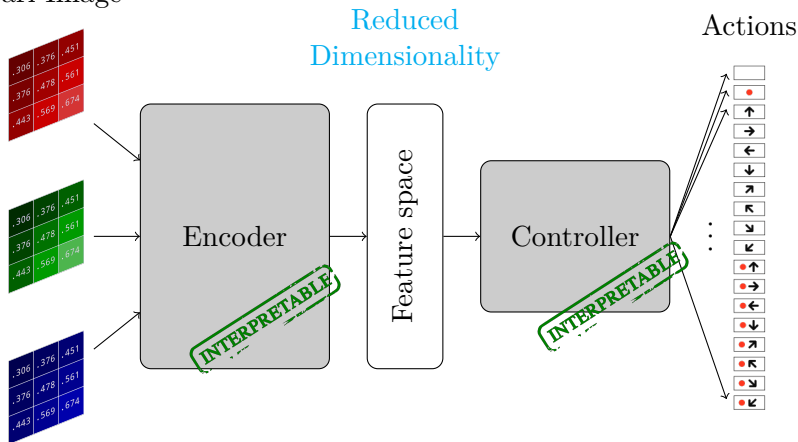
Approach: Interpretable Encoder – Controller

Atari Image



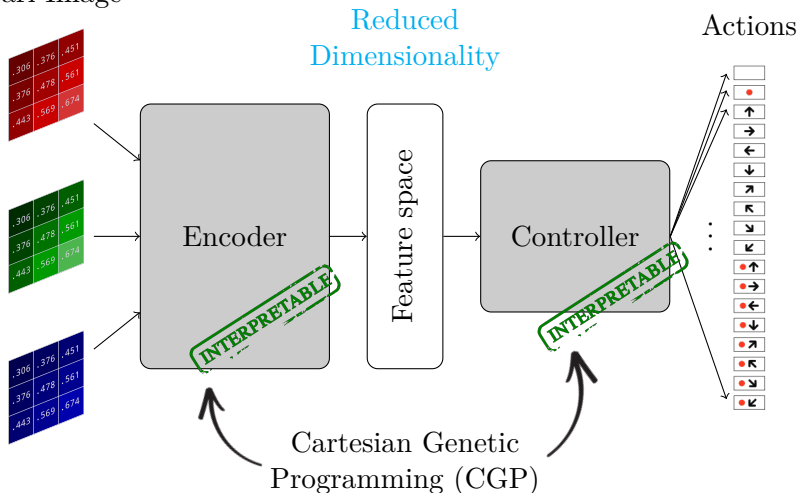
Approach: Interpretable Encoder – Controller

Atari Image



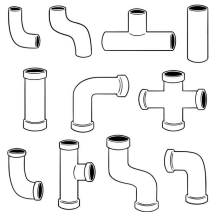
Approach: Interpretable Encoder – Controller

Atari Image

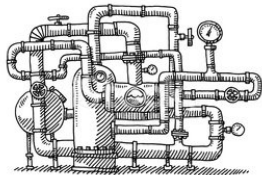


Cartesian Genetic Programming (CGP)

Cartesian Genetic Programming (CGP)



Evolutionary
Algorithm



Cartesian Genetic Programming (CGP)

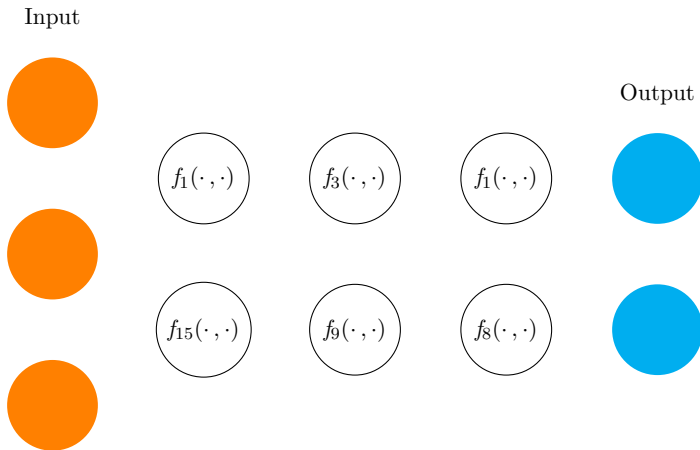
Input



Output

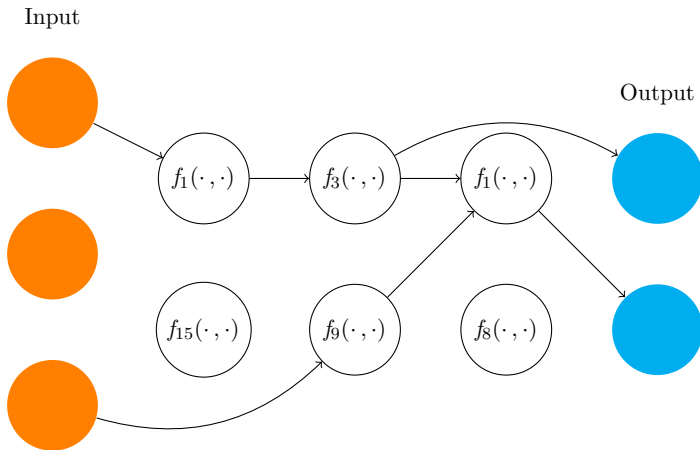


Cartesian Genetic Programming (CGP)



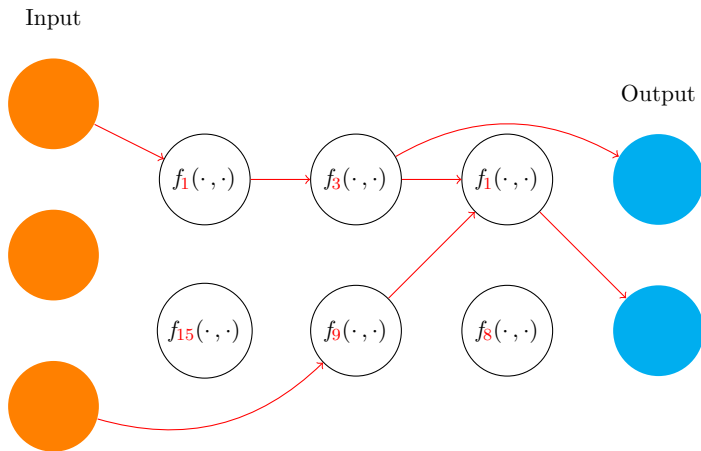
Function pool: $\{f_i : \mathcal{X}^2 \rightarrow \mathcal{X}\}_i$

Cartesian Genetic Programming (CGP)



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

Cartesian Genetic Programming (CGP)



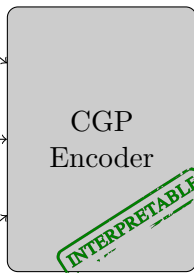
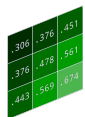
Function pool: $\{f_i : \mathcal{X}^2 \rightarrow \mathcal{X}\}_i$

Genotype:

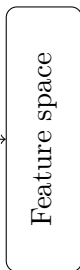
$[1, 1, 2, 13, 1, \dots, 3] \in \mathbb{N}^{3 \times \text{number of nodes} + \text{number of outputs}}$

Approach: Interpretable Encoder – Controller

Atari Image



Reduced
Dimensionality

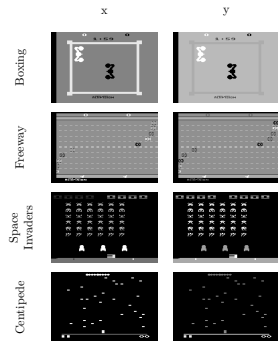


Actions

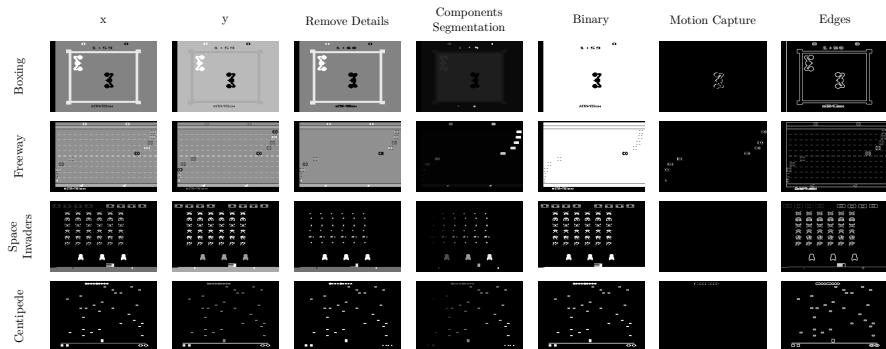


Encoder's Functions

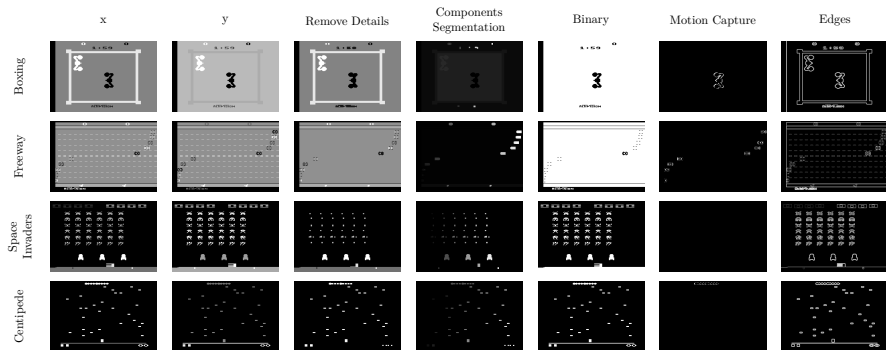
Encoder's Functions



Encoder's Functions



Encoder's Functions

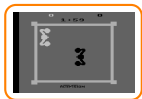
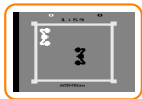


INTERPRETABLE

Encoder

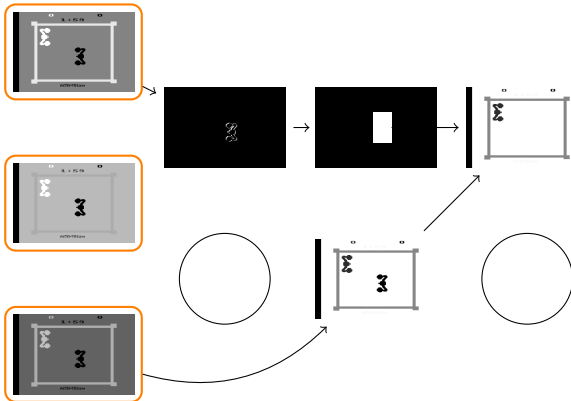
Encoder

Input

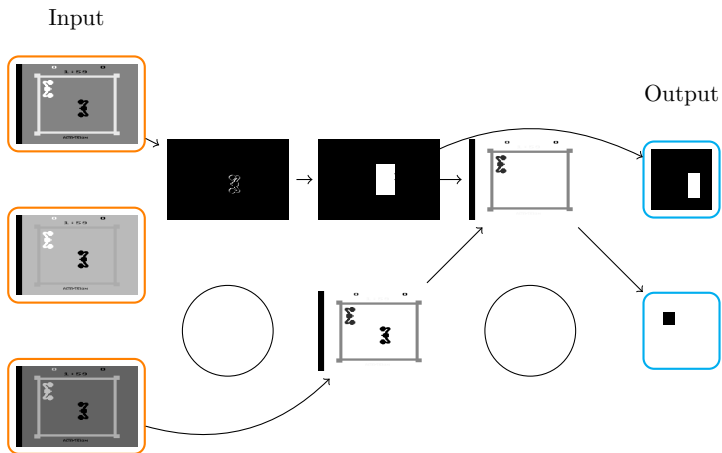


Encoder

Input

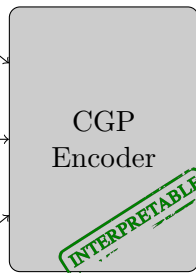
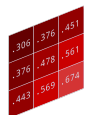


Encoder

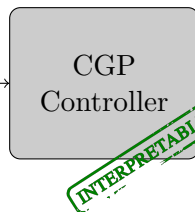
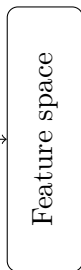


Approach: Interpretable Encoder – Controller

Atari Image



Reduced
Dimensionality

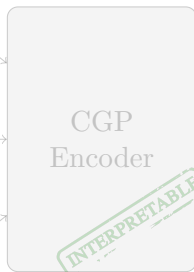
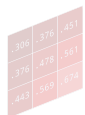


Actions



Approach: Interpretable Encoder – Controller

Atari Image



Reduced
Dimensionality



Actions



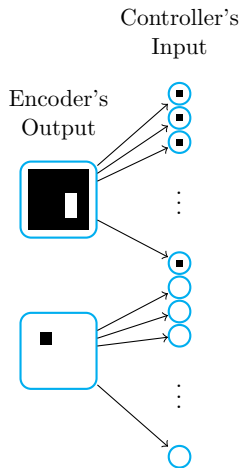
Controller

Controller

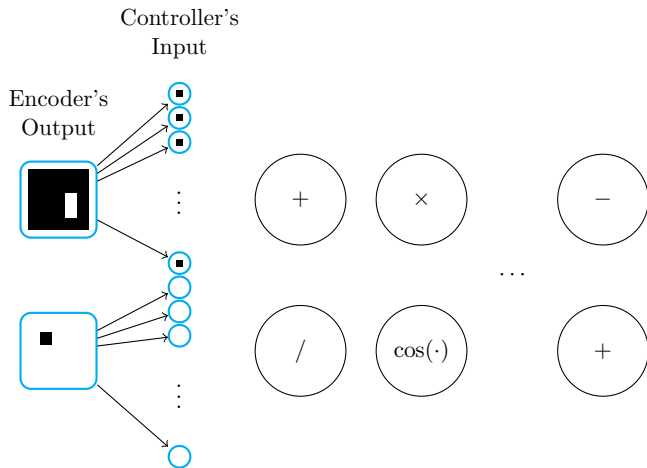
Encoder's
Output



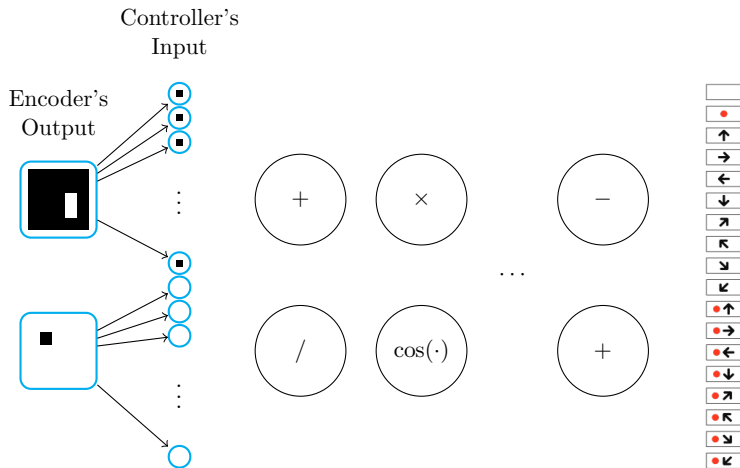
Controller



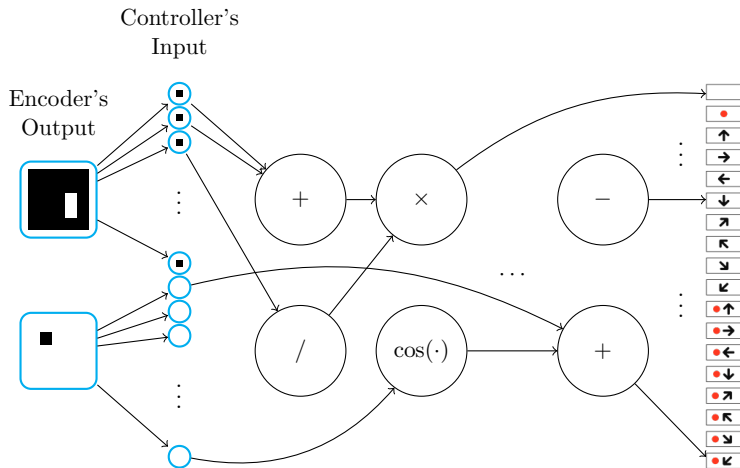
Controller



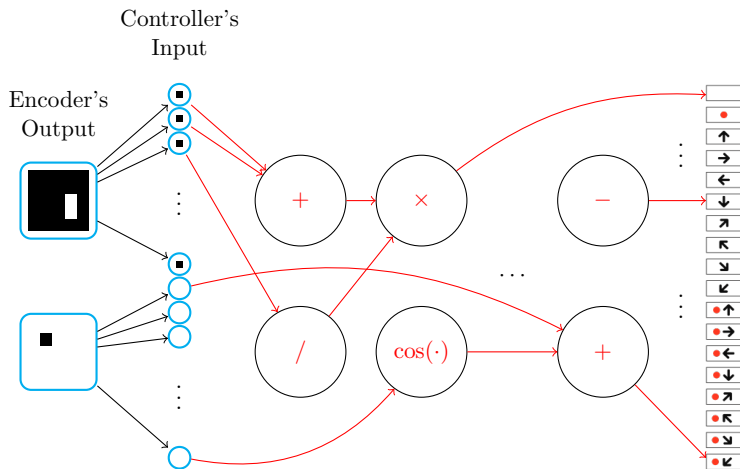
Controller



Controller



Controller

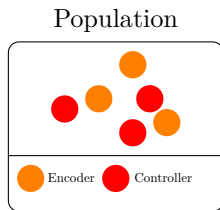


Genotype:

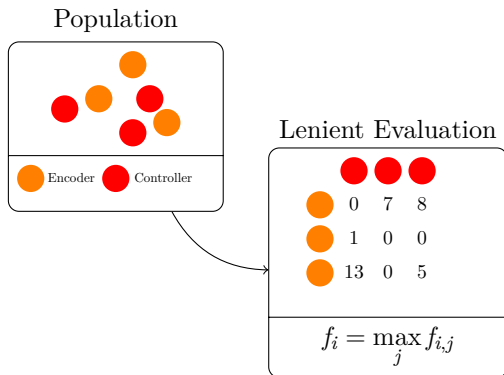
$[1, 1, 2, 13, 1, \dots, 3] \in \mathbb{N}^{3 \times \text{number of nodes} + \text{number of outputs}}$

Co-evolution

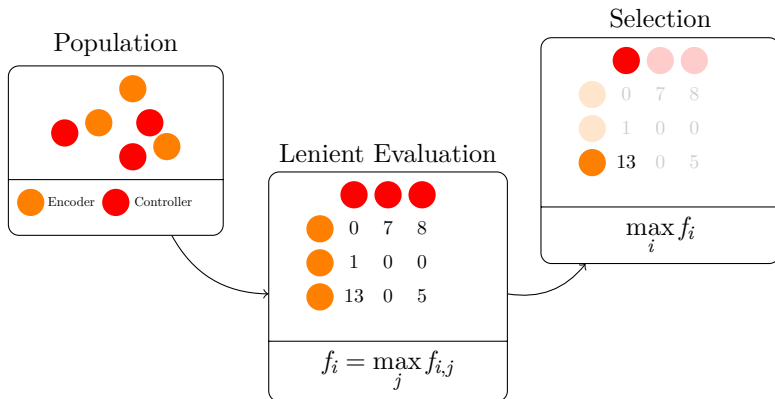
Co-evolution



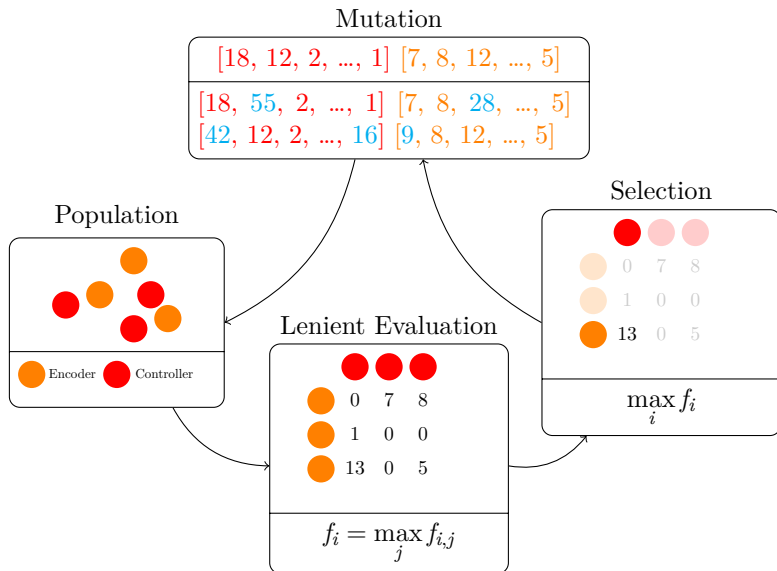
Co-evolution



Co-evolution



Co-evolution



Conclusion

Conclusion

- ▶ Objective: interpretable agent in pixel-based Atari games

Conclusion

- ▶ Objective: interpretable agent in pixel-based Atari games
- ▶ Approach: CGP co-evolution in an encoder-controller scheme

Conclusion

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

Conclusion

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

Conclusion

- ▶ Objective: interpretable agent in pixel-based Atari games
- ▶ Approach: CGP co-evolution in an encoder-controller scheme
- ▶ Encoder: interpretable image processing functions
- ▶ Controller: interpretable scalar functions
- ▶ Experiments: running

THANK YOU