ICML
International Conference On Machine Learning

Stabilizing Off-Policy Deep Reinforcement Learning from Pixels

Edoardo Cetin, Philip J. Ball*, Steve Roberts, Oya Celiktutan*



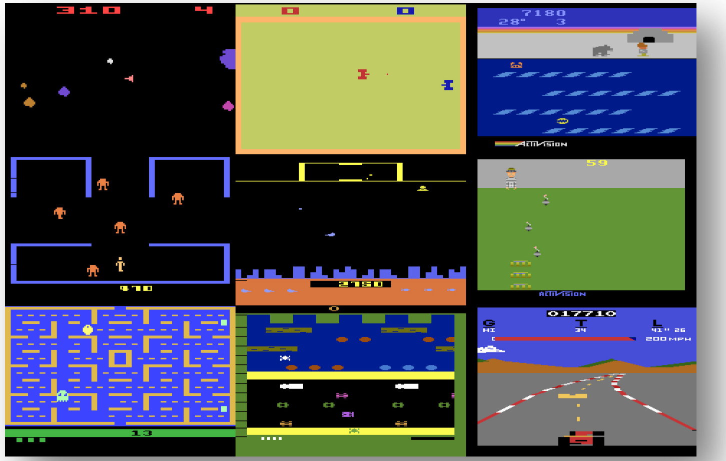

SAIR
Social AI & Robotics Lab
King's College London
University of Oxford

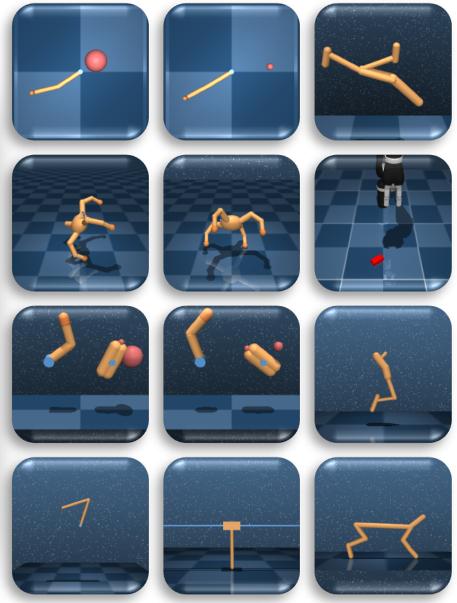
Instabilities in Pixel-Based RL

- Pixel observations pose concrete challenges for off-policy reinforcement learning (RL).

- Popular algorithms make use of several ***domain-specific*** practices.

  i.e. no ‘general purpose’ implementation for different benchmarks/problems.





Algorithm	Visual Deadly Triad Mitigation		
	TD-Loss	CNN Overfit	Low-Density Reward
DrQ/RAD	-	Shift/Jitter Augmentations	10-step returns†
DrQ-v2	-	Shift Augmentations	3-step returns
SAC-AE	VAE Loss	-	-
SPR	Model-Based Loss	Shift/Jitter Augmentations	10-step returns
DER	-	Non-Overlapping Strides	20-step returns
CURL	Contrastive Loss	Shift Augmentations	20-step returns*

The Visual Deadly Triad

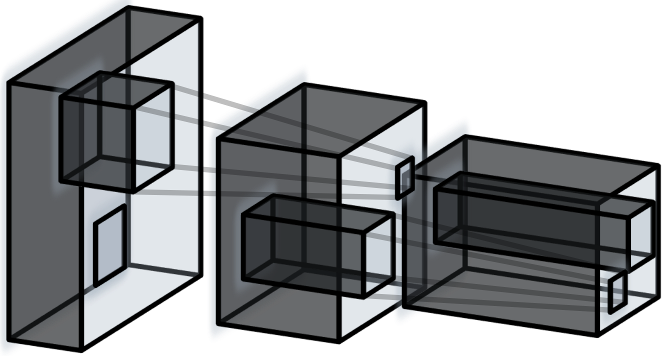
- We focus our empirical analysis on random shift augmentations on the DeepMind Control Suite.
  - Augmentations appear to counteract instabilities **exclusive** to off-policy critic learning.

- We observe these instabilities arise with the joint presence of three elements ***(visual deadly triad)***:

1. Learning the critic's weights solely from a temporal difference (TD) learning objective.
2. End-to-end backpropagation through unregularized convolutional encoders.
3. Low-magnitude, sparse environment rewards.

Agent	Final TD-Loss	Final Policy Loss	Return
Augmented	0.021	−0.99	86.5 ± 11.3
Non-Augmented	0.002	−1.05	9.2 ± 12.1
Proprioceptive	0.012	−1.14	79.1 ± 7.7
Frozen CNN (random)	0.023	−0.95	43.6 ± 20.2
Frozen CNN (pre-trained)	0.012	−0.99	77.6 ± 18.5
Non-Augmented (norm r)	18.616	3.86	38.6 ± 16.5
Non-Augmented (10-step returns)	0.003	−1.24	36.5 ± 20.3

$$\boxed{\nabla_\theta (Q_\theta(s,a) - (r + Q(s',a')))^2}$$



Effects on the Critic

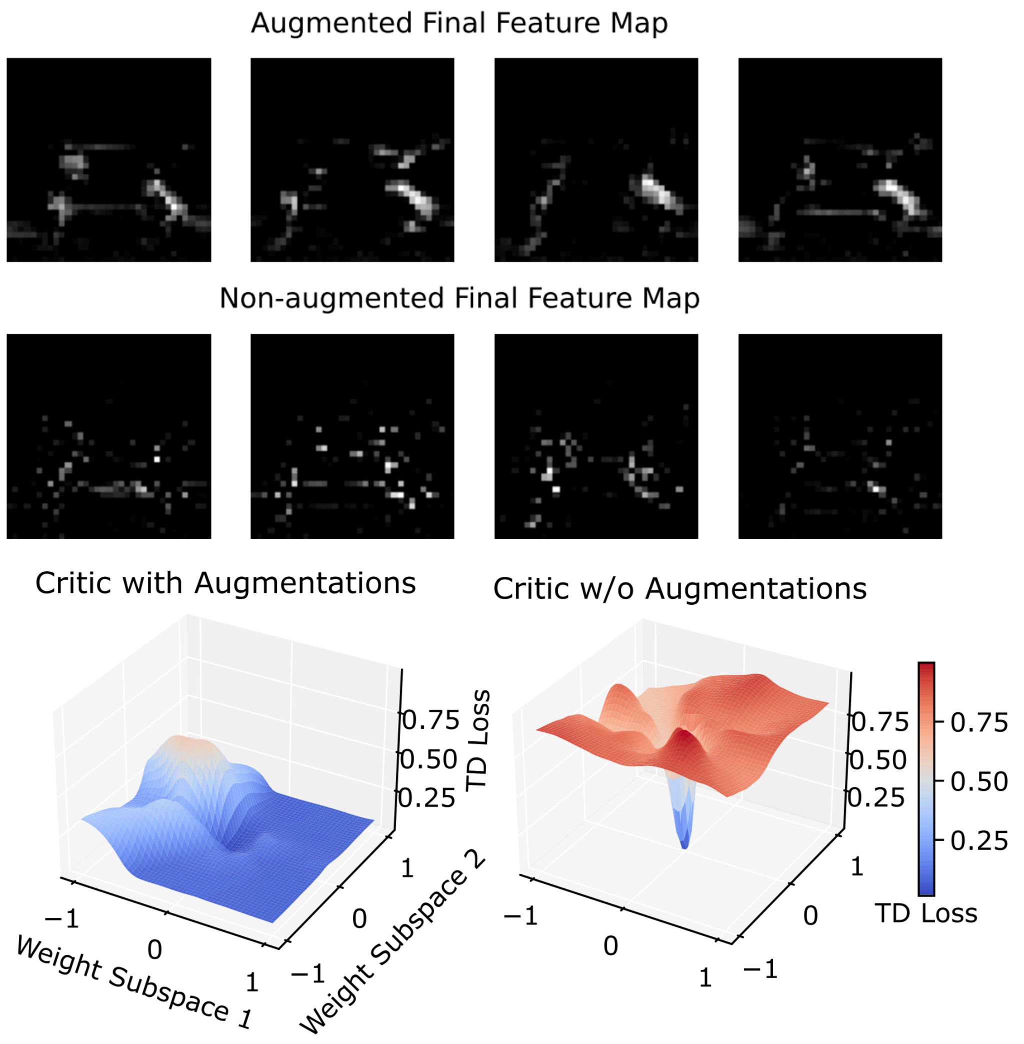
- Not addressing conditions in the visual deadly triad leads to ***catastrophic self-overfitting***

  i.e. encoder becomes susceptible to high-frequency noise, making the critic converges to a degenerate solution.

- We find this phenomenon can be measured from spatial discontinuities in the gradients via the Normalized Discontinuity (ND) score:

$$D(z)_{ijc} \approx \mathbb{E}_{v \sim S^1}\left[\left(\frac{\partial z_{ijc}}{\partial v}\right)^2\right],$$

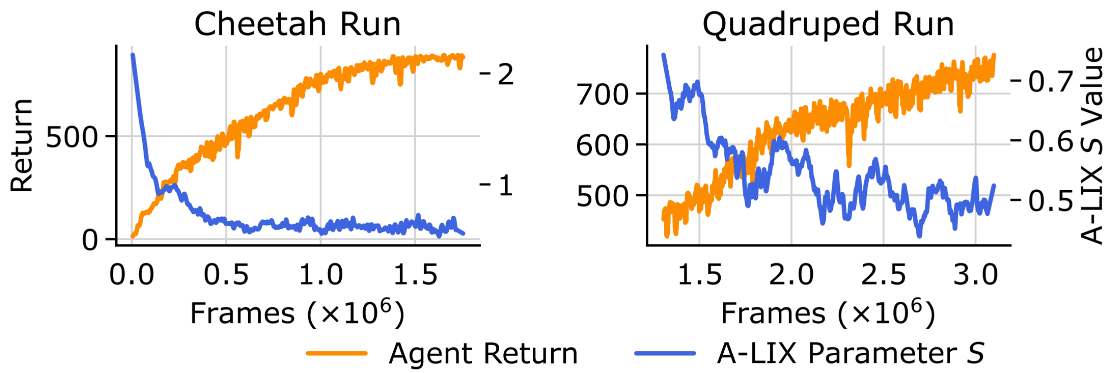
$$ND(z) = \frac{1}{C \times H \times W} \sum_{c=1}^{C} \sum_{j=1}^{H} \sum_{i=1}^{W} \frac{D(z)_{ijc}}{z_{ijc}^2}.$$

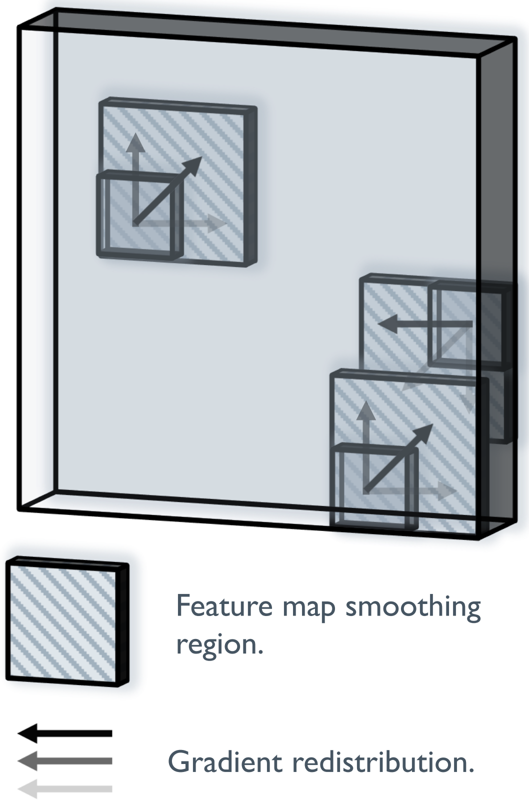


Encoder loss surface plots.

Adaptive Local Signal Mixing (A-LIX)

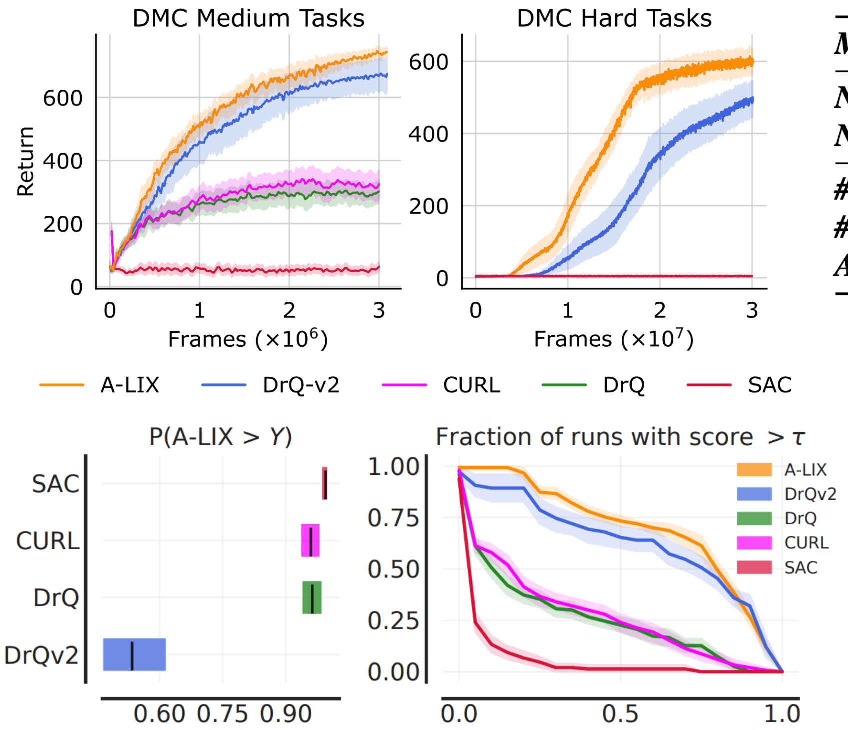
- To prevent this phenomenon we design **A**daptive **L**ocal **S**ignal **M**ixing ***(A-LIX)***:
  - Smoothing the features in a random local neighborhood within each feature map.
  - During backpropagation discontinuous gradients get randomly redistributed.
- We adaptively tune the magnitude of the regularization ***(S)*** based on the ND scores.





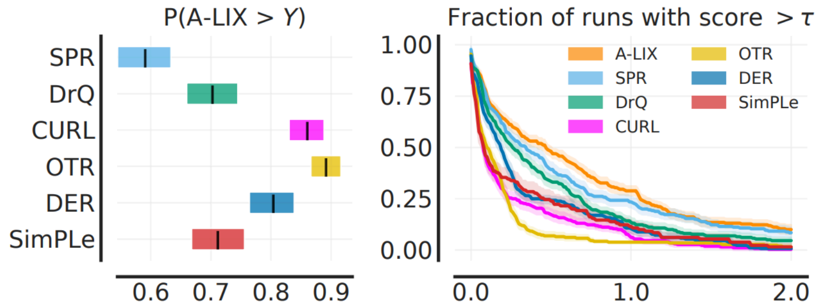
Performance Results

A-LIX outperforms SOTA off-policy baselines w/o many auxiliary practices and the same network architecture for both DMC and Atari 100k.



(a) DeepMind Control: Medium and Hard Tasks

Metrics	SimPLe	DER	OTRainbow	CURL	DrQ	SPR	**A-LIX**
Norm. Mean	0.443	0.285	0.264	0.381	0.357	0.704	**0.753**
Norm. Median	0.144	0.161	0.204	0.175	0.268	**0.415**	0.411
# SOTA	7	1	1	1	1	4	**11**
# Super	2	2	1	2	2	**7**	**7**
Average Rank	3.92	5.00	5.21	3.92	4.85	2.88	**2.21**



(b) Atari 100k

Conclusion

Our work's contributions for pixel-based off-policy RL:

- A new hypothesis *to explain* instabilities.
- A new score *to detect* instabilities.
- A-LIX, a new regularization layer *to prevent* instabilities.

For further details and access to our open-source code, please visit:

sites.google.com/view/a-lix/home

SCAN ME

Thank you, and see you at the poster session.