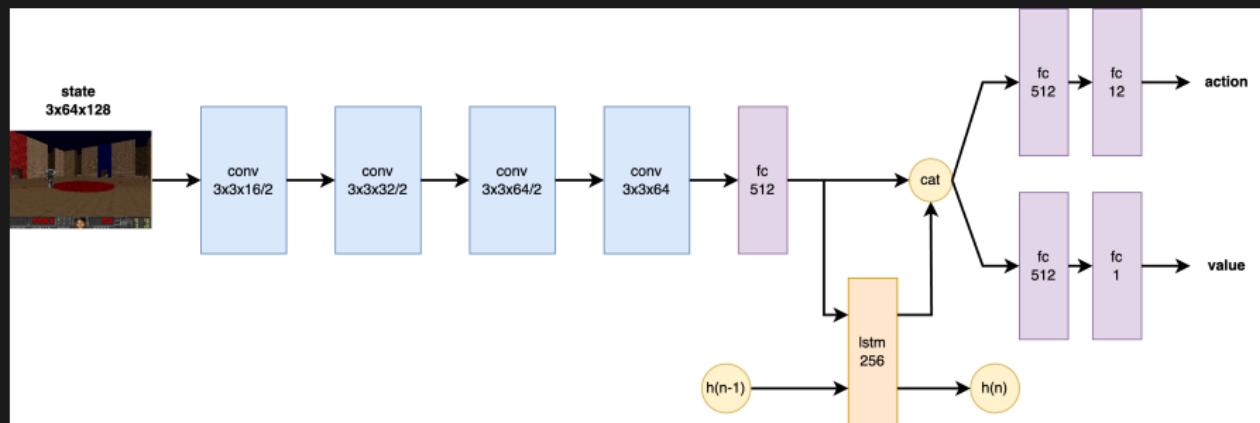


reinforcement learning with self supervised learning

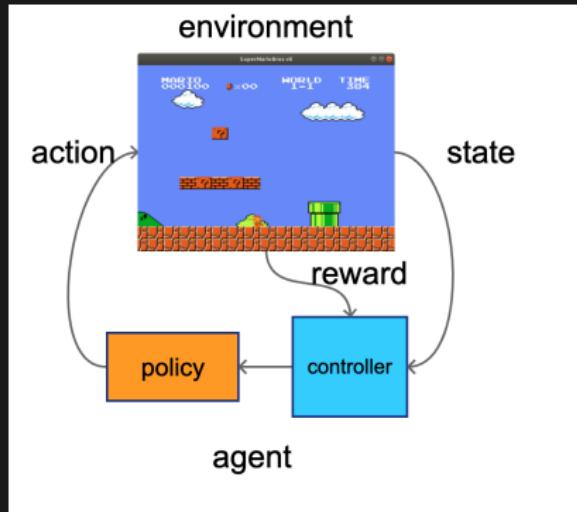
- Conquista of Montezuma's Revenge

Michal CHOVANEC, PhD.

net of doom - doom playing model



reinforcement learning



- ① **obtain state** - observation
- ② **choose action** - policy
- ③ **receive reward**
- ④ **learn from experiences**

reinforcement learning - success story

Mastering the game of Go with deep neural networks and tree search, Nature, 2016



reinforcement learning - success story

RMA: Rapid Motor Adaptation for Legged Robots

A) Training in Simulation

Phase 1

Mass, COM, Friction
Terrain Height
Motor Strength
(e_t)

x_t, a_{t-1}

Env Factor Encoder (μ)

Z_t

Base Policy (π)

*Trainable Modules in Red

Phase 2

x_{t-51}, a_{t-51}

:

x_{t-1}, a_{t-1}

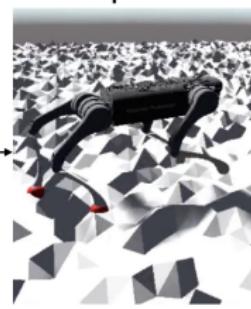
x_t, a_{t-1}

Adaptation Module (ϕ)

Regress

\hat{Z}_t

Base Policy (π)



Physics Simulation

B) Deployment

x_{t-50}, a_{t-51}

:

x_t, a_{t-1}

x_t, a_{t-1}

Adaptation Module (ϕ)

10 Hz

\hat{Z}_t

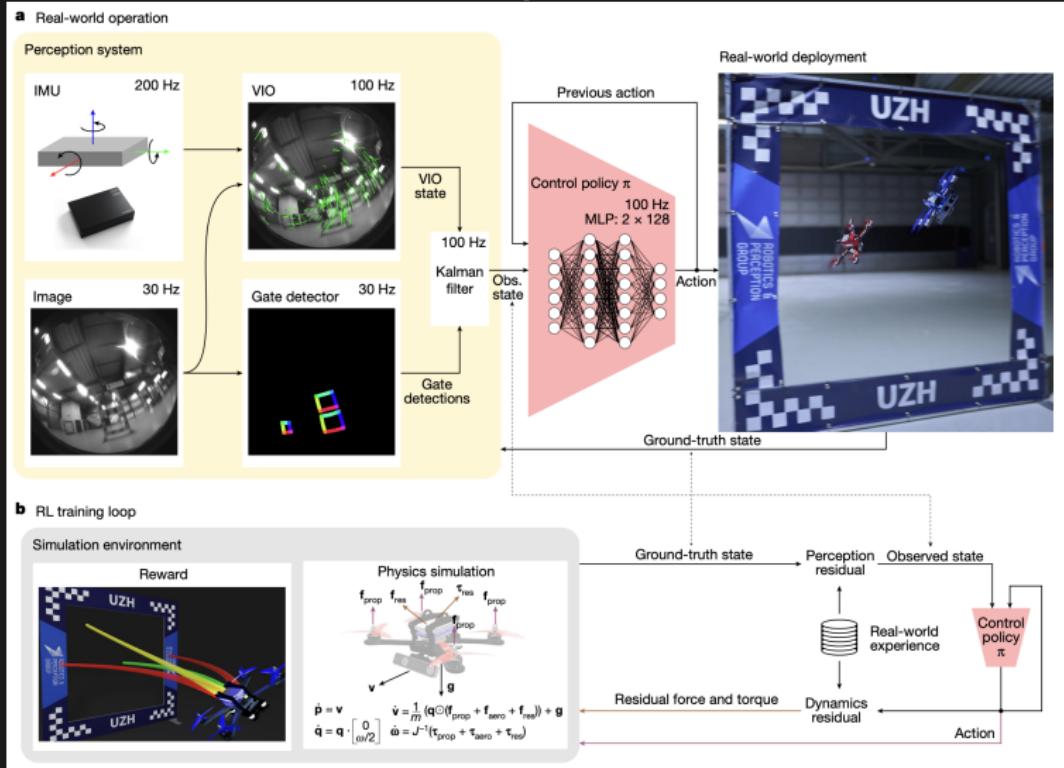
Base Policy (π)

100 Hz



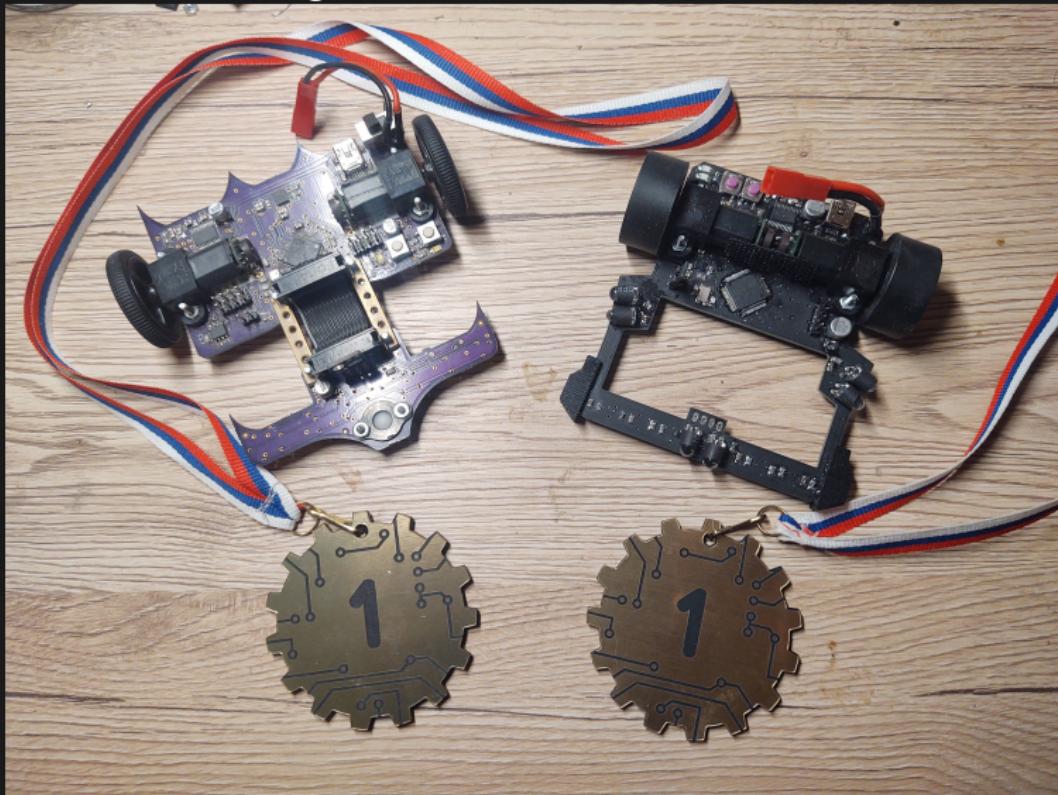
reinforcement learning - success story

Champion-level drone racing using deep reinforcement learning, 2021



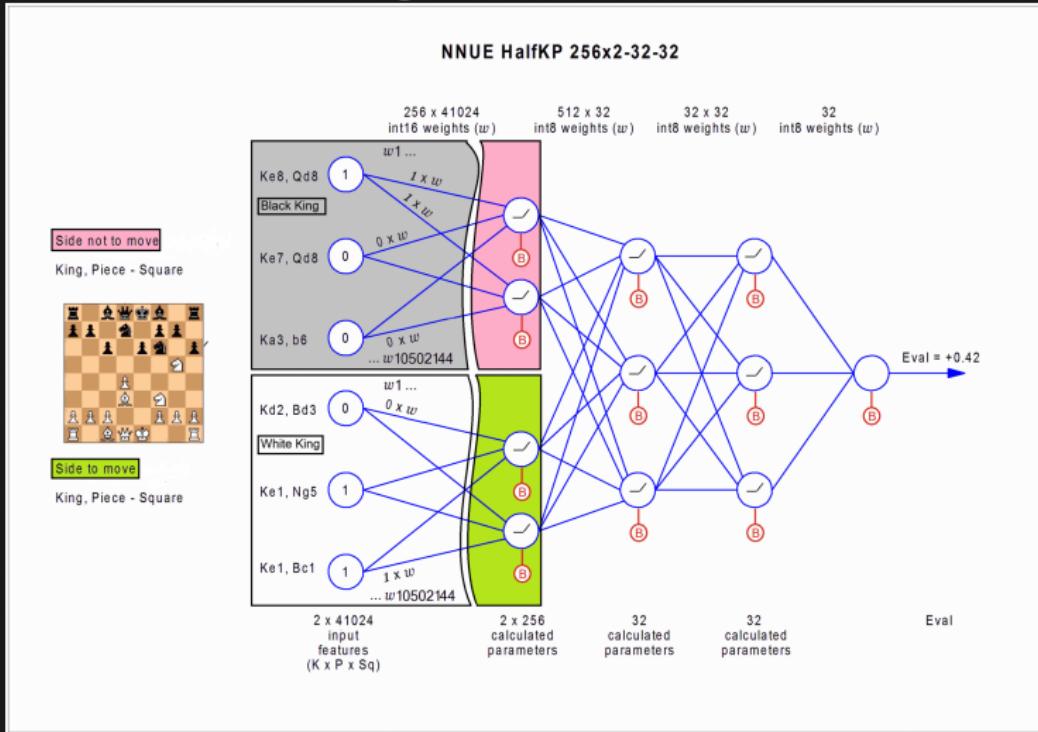
reinforcement learning - success story

Motoko line following robot

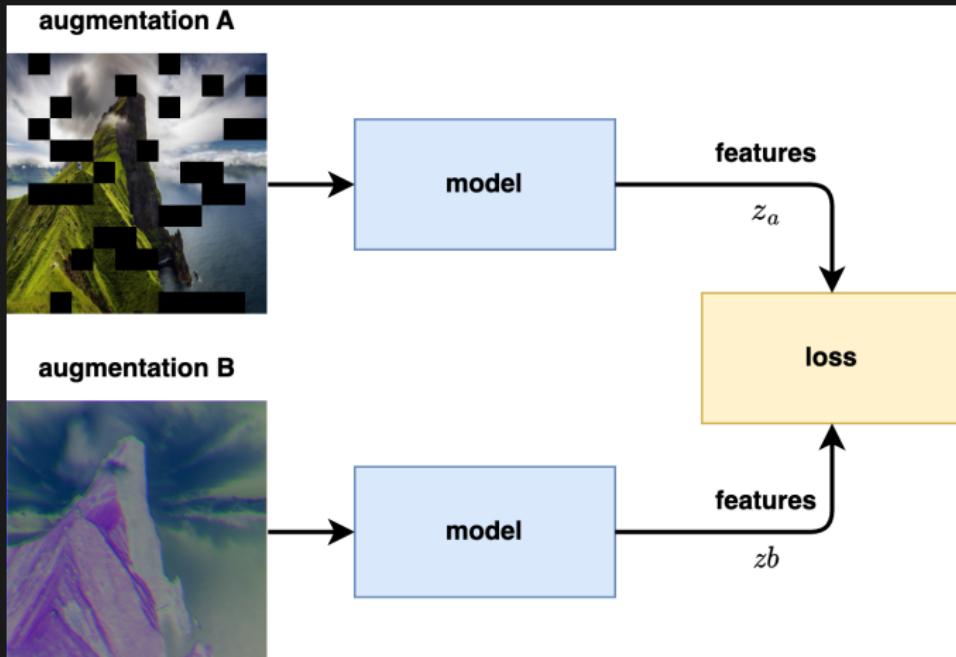


reinforcement learning - success story

Stockfish NNUE Chess engine

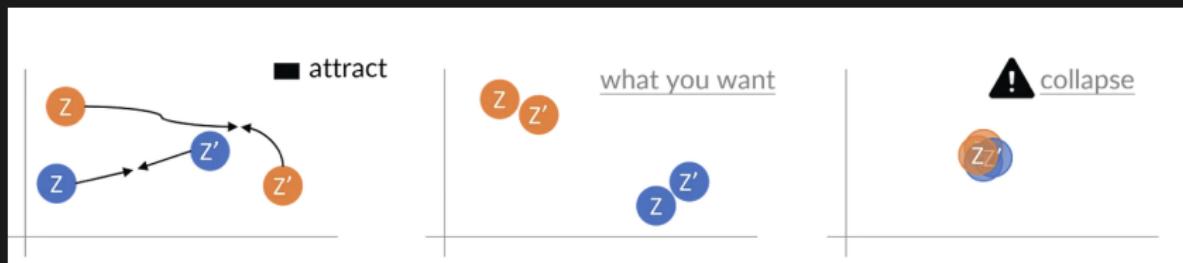


self supervised learning

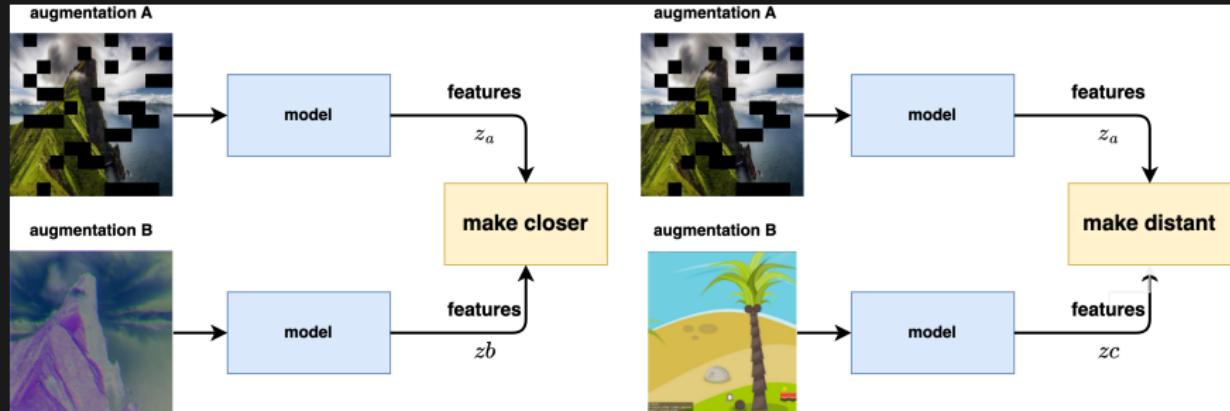


$$\mathcal{L} = \|z_a - z_b\|_2^2$$

self supervised learning - loss collapse



self supervised learning - contrastive loss



VICReg - non contrastive supervised learning

- VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning
- Yann LeCun: Dark Matter of Intelligence and Self-Supervised Learning



VICReg - non contrastive self supervised learning

- VAriance

$$\mathcal{L}_{variance} = \frac{1}{N} \sum_n \max(0, 1 - std(z_{n,:}^*))$$

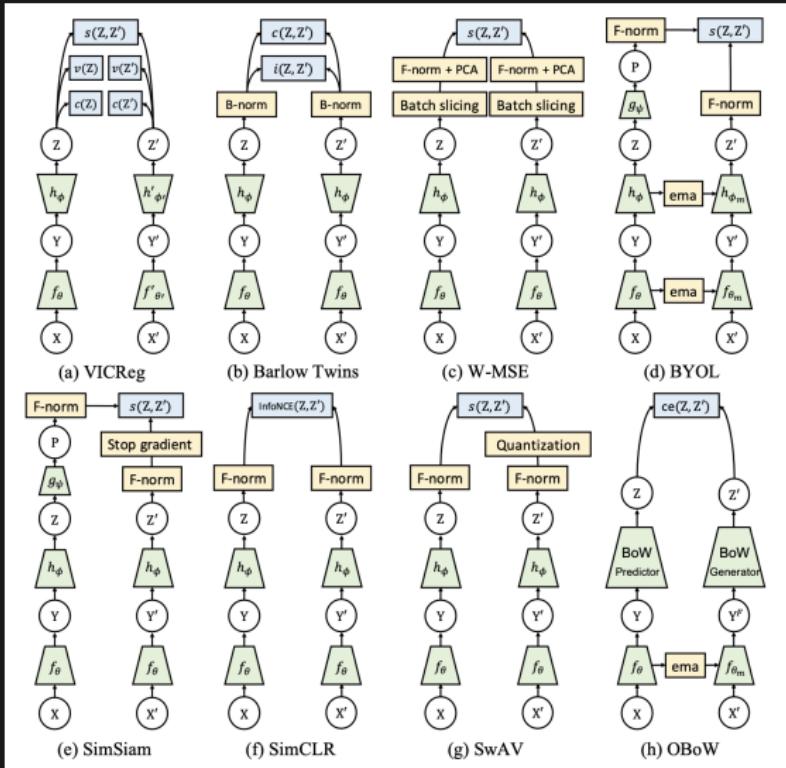
- INvariance

$$\mathcal{L}_{invariance} = \frac{1}{N} \sum_n |z_{n,:}^a - z_{n,:}^b|_2^2$$

- Covariance

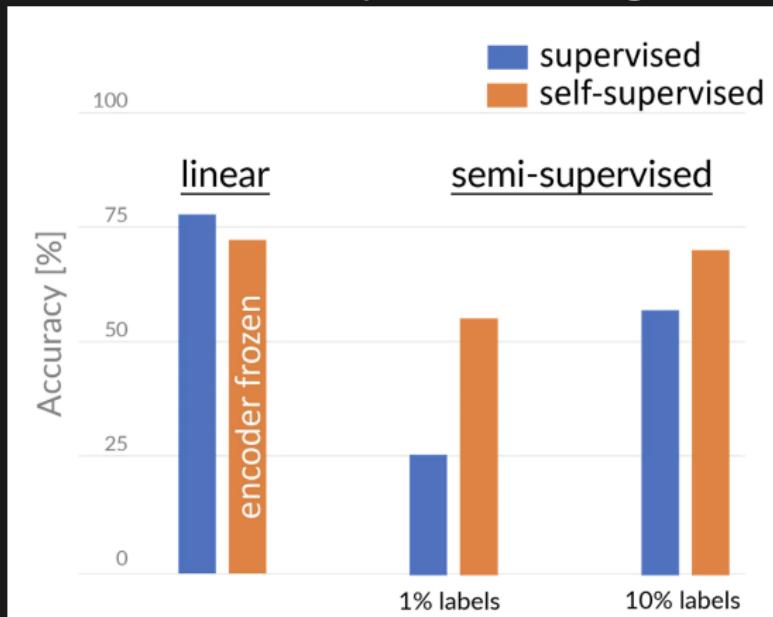
$$\mathcal{L}_{covariance} = \frac{1}{F} \sum_{i \neq j} cov(z_{:,i}^*, z_{:,j}^*)$$

another loss examples



applications

- face, signature matching
- anomaly detection
- e-shop similar goods finding
- searching by image - Google lens
- less labeled data for self supervised training



- I-JEPA : The first AI model based on Yann LeCun's vision for more human-like AI
- MAST : Masked Augmentation Subspace Training for Generalizable Self-Supervised Priors

Montezuma's revenge - 10 years of tears?

source : <https://paperswithcode.com/sota/atari-games-on-atari-2600-montezumas-revenge>

year	name	score
2013	Playing Atari with Deep Reinforcement Learning	0
2015	Deep Reinforcement Learning with Double Q-learning	0
2017	Curiosity-driven Exploration by Self-supervised Prediction ^a	0
2021	MuZero	2500
2018	Count-Based Exploration with Neural Density Models ^b	3705
2019	Exploration by Random Network Distillation ^c	8152
2021	GoExplore* ^d	43 000

* requires environment state saving/loading

^a<https://arxiv.org/abs/1705.05363>

^b<https://arxiv.org/abs/1703.01310>

^c<https://arxiv.org/abs/1810.12894>

^d<https://arxiv.org/abs/2004.12919>

sample efficiency

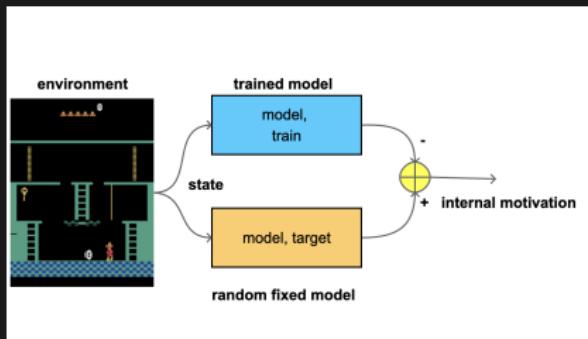
- RND ¹ $4.5 * 10^9$ samples, score 8152 on MR
- Never give up ² $3.5 * 10^{10}$ samples, score 10 000 on MR
- SND $1.28 * 10^8$ samples with score 25 000

NGU on my machine means **740 days !!!**

¹Burda et al. 2018

²Badia et al. 2020

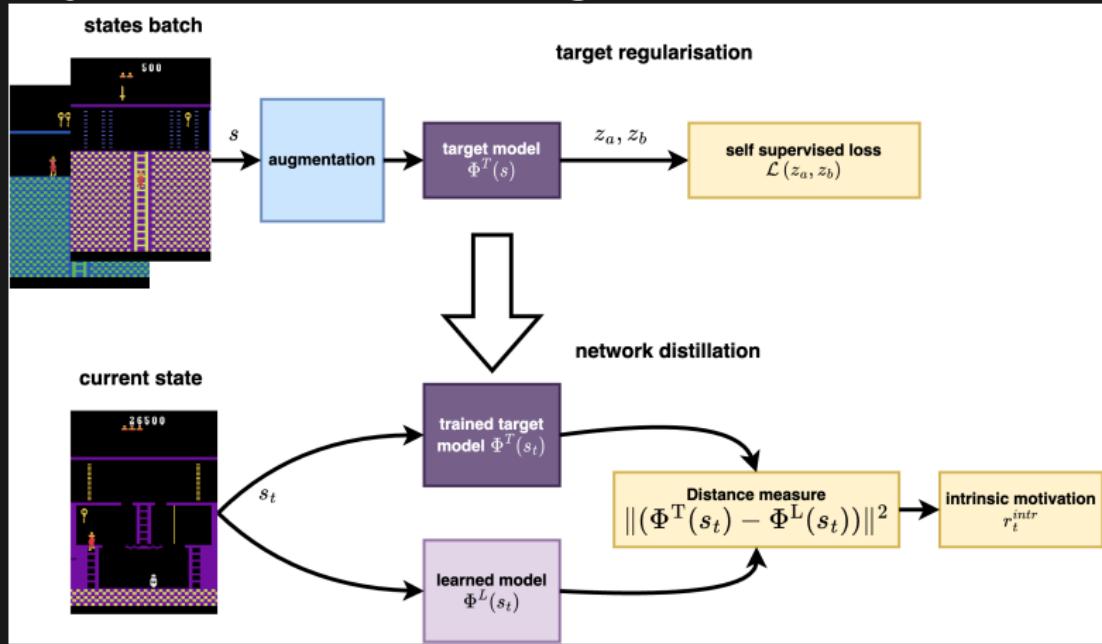
random network distillation



- neural network works as **novelty detector**
- model learns to imitate random (target) model
- **less visited states produce bigger motivation signal**
- orthogonal weights
initialisation ($g = 2^{0.5}$) for strong signal
- lot of fully connected layers
to avoid generalisation

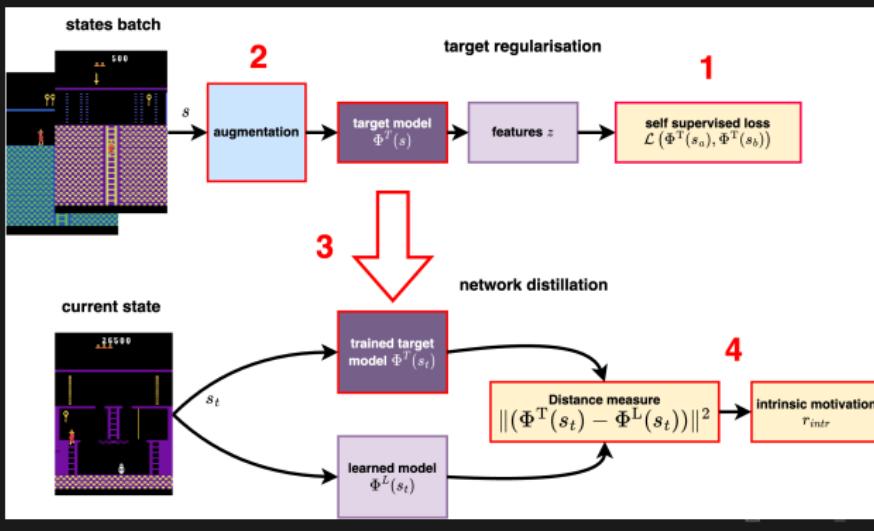
Exploration by self supervised-exploitation

Matej Pecháč, Michal Chovanec, Igor Farkaš



Exploration by self supervised-exploitation

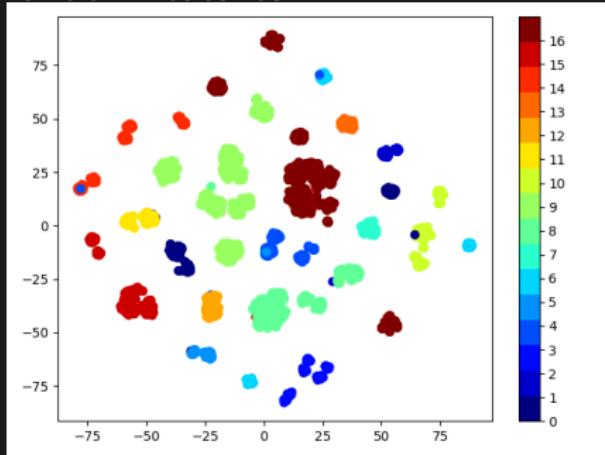
- we extended existing idea of Random Network Distillation
- 1 : for target model, self supervised training is used
- 2 : augmented states are used to train target model
- 3 : target model is used as distillation source
- 4 : distillation error is used for intrinsic motivation



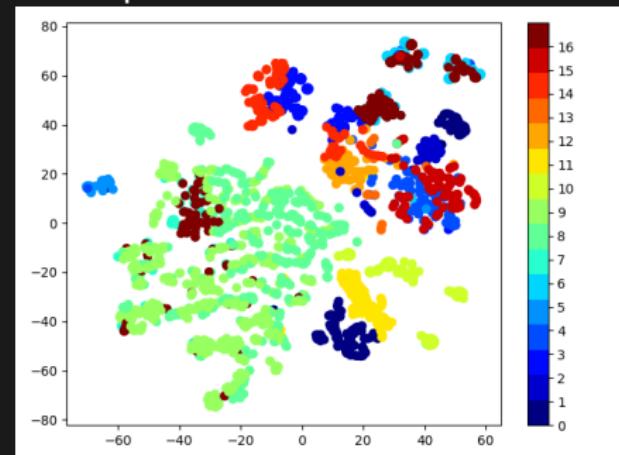
Trained features

- t-SNE features projection for random and trained models
- color represents different rooms in Atari Montezuma's Revenge
- self supervised features provides much bigger variance
- preventing agent to stuck

random features



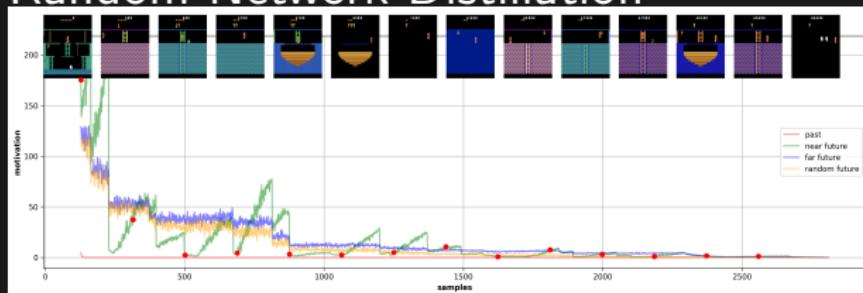
self supervised trained features



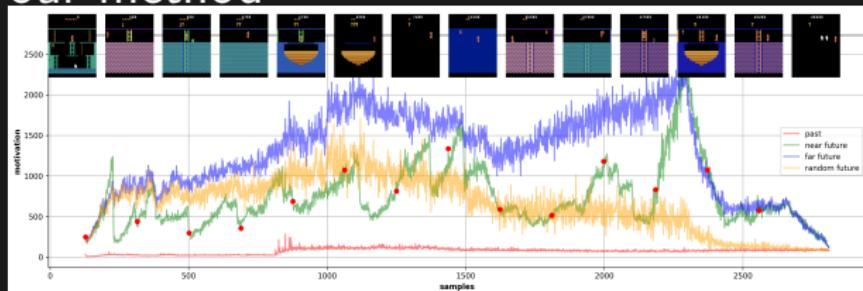
Exploration signal

- Random Network Distillation signal decrease over time
- our method provides more informative signal

Random Network Distillation

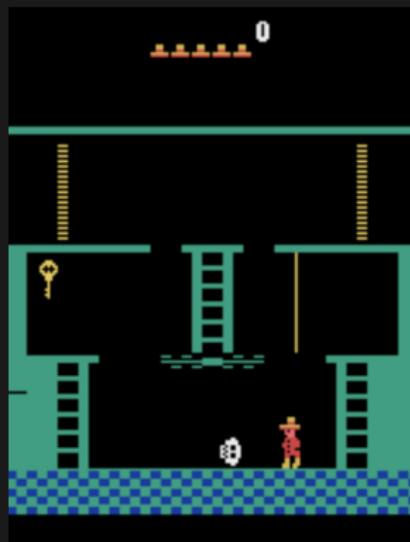


our method



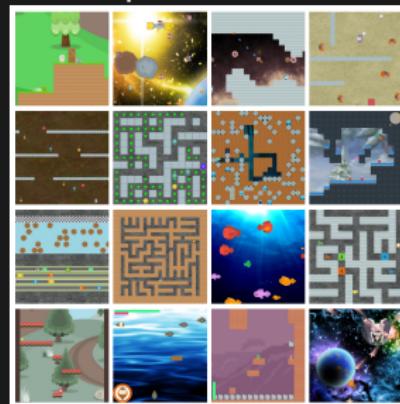
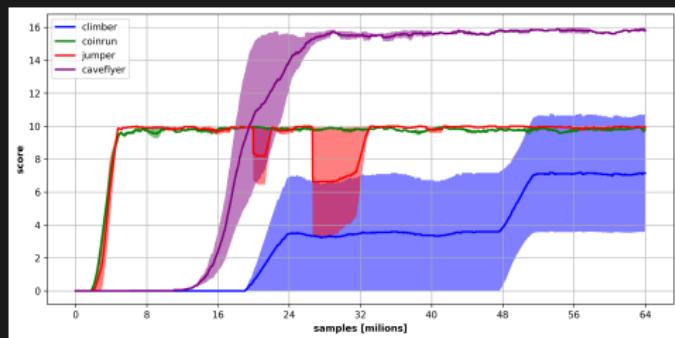
Results

- Montezuma's Revenge, with score 25 000+
- Private Eye, with score 12 000+
- Venture, Gravitar
- 128M samples total - only single GPU needed



Results

solved Procgen hard exploration seeds environments : Caveflyer, Climber, Coinrun, Jumper



misleading papers - Curiosity-driven Exploration by Self-supervised Prediction ^a

^aPathak et al. 2017

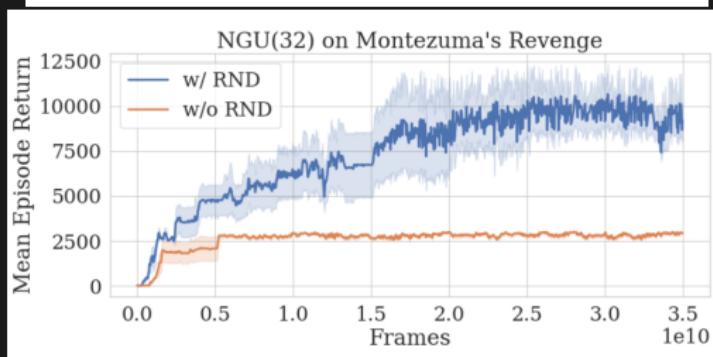
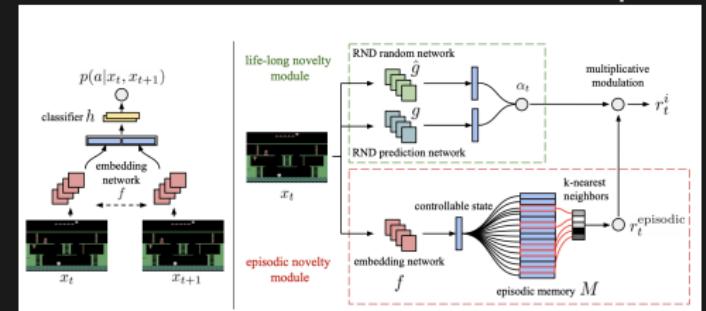
- not working on REAL hard exploration problems
- Super Mario is special case - moving forward is close to optimal policy
- inverse ICM model - why they didn't show accuracy (my results around only 40% !!!)
- how predicted state looks ?

	Gravitar	Montezuma's Revenge	Pitfall!	PrivateEye	Solaris	Venture
RND	3,906	8,152	-3	8,666	3,282	1,859
PPO	3,426	2,497	0	105	3,387	0
Dynamics	3,371	400	0	33	3,246	1,712
SOTA	2,209 ¹	3,700 ²	0	15,806²	12,380¹	1,813³
Avg. Human	3,351	4,753	6,464	69,571	12,327	1,188

misleading papers - Never give up ^a

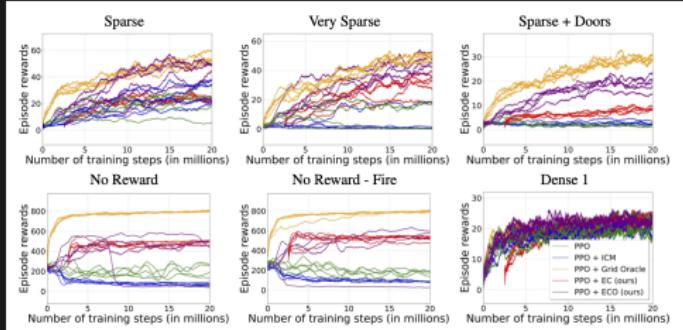
^aBadia et al. 2020

nice looking score, but on cost of $3.5 * 10^{10}$ samples !!!



other misleadings

avoiding comparing with SOTA or common benchmarks
results : Episodic Curiosity Through Reachability, Savinov, 2019



many other :

- simple gridworld or toy environment experiments
- providing key prior information (e.g. position)
- selecting only "good" results

Q&A



- <https://github.com/michalnand/>
- michal.nand@gmail.com