Incorporating Uncertainty into Decision Making

CPS: Small: Informed Contextual Bandits to Support Decision-Making for Intelligent CPS

Daniel Krutz (PI), Travis Desell (Co-PI), Alexander Ororbia (Co-PI). Rochester Insitute of Technology

Can We Ignore Labels in Out-of-Distribution Detection?

Problem: Safety-critical AI systems need to reliably identify out-of-distribution (OOD) inputs, which have labels not seen during training, but many current approaches try to do this without labels

Key Finding: Label Blindness The paper introduces the concept of "label blindness," proving that self-supervised and unsupervised methods will fail at OOD detection when their learning objective is independent of the information needed to predict labels

- This means that if the method does not learn about the labels during training, it cannot detect when a new input does not match those labels.

Implications: It is not safe to ignore labels for OOD detection in real-world applications, as methods trained without labels may fail if the task depends on features that are independent of the training process. Future research should focus on incorporating label information, even if only a small amount, to improve OOD detection.

The paper theoretically proves that when there is zero mutual information between the learning objective and the in-distribution labels, OOD detection will fail.

The paper demonstrates that it is not safe to ignore labels for OOD detection, contrary to the recent trend of unlabeled OOD methods.

The Label Blindness Theorem indicates that there cannot be a single unlabeled OOD detection algorithm applicable to all unlabeled data.

Future work should focus on methods that incorporate label information, such as few-shot or one-shot methods.

Hong Yang, Qi Yu, Travis Desell. **Can We Ignore Labels in Out-of-Distribution Detection?** *The Thirteenth International Conference on Learning Representations (ICLR 2025)*. Singapore. April 24-28, 2025.

An example failure case by visualizing the heatmaps of the gradient of a unlabeled SimCLR trained Resnet. The OOD detection task is to detect OOD facial expressions.

In this case, the OOD detection method fails as justified by our theoretical work, where the representations do not exhibit a strong gradient in regions commonly associated with facial expressions (i.e., eyebrows, mouth, etc.)

Table 1: Results from experiments across various datasets and methods. Unlabeled methods perform poorly in adjacent OOD detection. CLIPN performance is due to labels present in the pretraining dataset and is discussed in section 5.3. Higher AUROC and lower FPR is better.

	Faces		Cars		Food	
Method	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95
Supervised MSP	70.8±0.3	88.2±0.2	69.2±0.9	88.8±0.8	78.8±1.2	81.1±1.6
SimCLR KNN	52.0±4.2	95.0±1.3	52.5±0.4	94.0±0.5	61.1±2.8	91.6±1.6
SimCLR SSD	55.0±4.5	95.1±2.0	52.7±0.7	93.7±1.1	64.4±0.8	89.3±0.5
RotLoss KNN	46.1±2.5	95.8±0.4	51.1±0.6	94.8±0.7	49.7±3.8	94.9±0.9
RotLoss SSD	46.6±3.0	95.7±0.5	50.7±1.9	95.0±1.2	50.7±3.6	94.9±0.9
Diffusion LPIPS	54.7±4.6	94.2±3.7	53.8±1.8	93.9±1.2	52.9±2.2	94.4±0.6
Diffusion MSE	55.3±2.2	94.2±1.4	51.6±1.6	94.4±0.5	52.5±3.4	94.2±0.6
CLIPN CTW	47.0±1.4	97.3±0.3	65.0±5.1	69.4±9.4	70.9±2.9	69.1±7.0
CLIPN ATD	44.2±1.4	97.5±0.2	81.1±4.3	56.6±10.4	84.9±0.2	53.9±4.5
CLIPN MSP	58.7±4.4	95.9±1.4	76.5±1.4	75.4±0.6	80.5±1.6	74.0±1.4

Uncertainty-Aware Reinforcement Learning Through Uncertainty-Driven Replay Memory (UDRM)

- Novel idea
- Populate the replay buffer, with experiences, using the aleatoric and epistemic uncertainties
- Compared performance in 5 MinAtar environments against existing models for 10 seeds each and 2.5M time steps
  - Uncertainty-Aware Deep Q Network (UADQN)
  - Calibrated Evidential Quantile Regression in Deep in Network (CEQR-DQN)
  - Proportional Prioritized Experience Replay (PER-Prop)
- The current code promotes more exploitation

	UADQN	CEQR-DQN	CEQR-DQN with PER	UDRM (α=50)	UDRM (α=500)
CartPole	**88.4499 ± 12.8856**	39.141 ± 24.4919	43.1785 ± 7.618	50.3079 ± 7.1925	50.3584 ± 7.2659
Asterix	15.0796 ± 1.4093	34.1909 ± 3.7789	27.1338 ± 3.804	32.8226 ± 3.6796	**36.0932 ± 3.294**
Breakout	15.2167 ± 2.2581	13.4228 ± 4.0635	10.7632 ± 5.2481	14.5814 ± 5.5391	**17.6408 ± 7.2645**
Freeway	**56.7454 ± 1.6765**	51.3018 ± 1.8275	46.3032 ± 1.8215	51.4597 ± 2.0143	51.5763 ± 1.6764
Seaquest	8.0099 ± 4.6142	19.9233 ± 4.6946	18.3091 ± 4.7537	**22.6576 ± 5.9426**	22.0925 ± 4.1124
SpaceInvaders	78.4135 ± 4.4921	222.3827 ± 9.7841	**239.0748 ± 17.011**	220.3821 ± 17.0678	221.2928 ± 15.2414

UDRM performs the better than the other models for 3 out of the 6 games. Considering the resources required for the models, UDRM performs better than or at par with CEQR-DQN. The proportional priority experience replay buffer outperforms UDRM in 1 of the 6 games.

- Transitions with lower uncertainty are added more than once to the replay buffer to skew the distribution in favor of certain actions
- The agent explores without any hindrance for the first 10% of the time steps. During which the uncertainties, for each transition, are stored to calculate the initial alpha and beta values
- The transition uncertainties are compared against a threshold that is calculated as a function of the time step t. α and β are recalibrated at regular intervals.

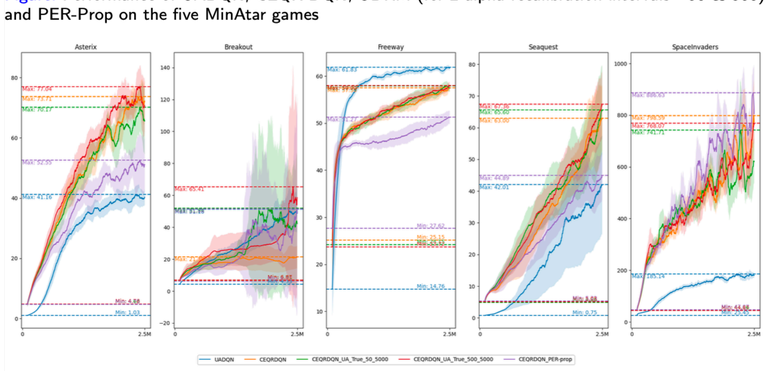
$$threshold = \alpha e^{-\beta t}$$

- The initial α is set as the 99th percentile of the uncertainties stored.
- The initial beta value is calculated such that it halves by the end of training, i.e. $k = 2$.

$$\beta_{initial} = \frac{ln(k)}{timesteps}$$

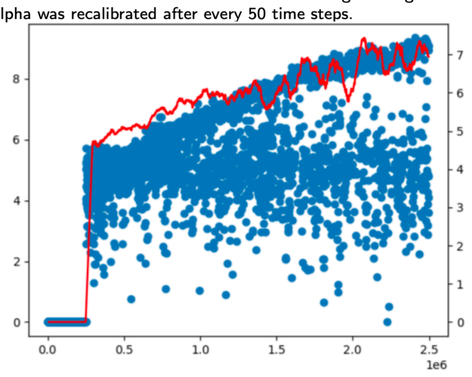
- The proposed UDRM model is tested in 6 game environments.

Figure: Performance of UADQN, CEQR-DQN, UDRM (for 2 alpha recalibration intervals - 50 & 500) and PER-Prop on the five MinAtar games



UDRM achieves a higher maximum score compared to UADQN, CEQR-DQN and CEQR-DQN with PER in the Asterix, Breakout and Seaquest environments. For most of the games, UDRM is able to achieve higher scores at each time step compared to the other models.

Figure: Percentage of actions that were remembered twice during training in the MinAtar Asterix environment and alpha was recalibrated after every 50 time steps.



As the training progresses, the model is able to reduce its uncertainty, below the threshold, for a greater number of experiences.

Informed Contextual Multi-Armed Bandits

A new framework to utilize forecasting methods or neural networks to serve below goals,

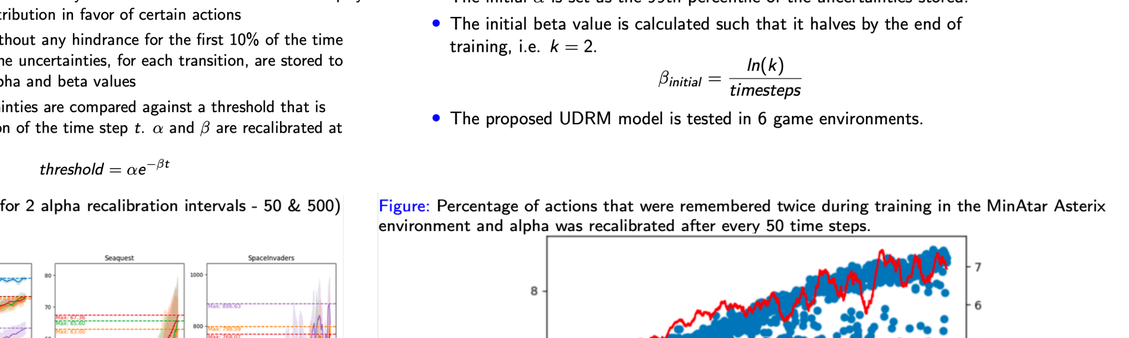
1. **Provide better decisions/system actions through**
   a. identifying corrupt context and/ or fulfilling missing context
   b. predicting future context
2. **Provide a measure of confidence in its predictions/ recommendations by addressing volatility and uncertainty**
3. **Drive real-time decisions in CPS**

Use Case: Stock Trading

- *Experimented with stock data for 30 companies.*
- *Adversarial bandit algorithm **EXP4.P** with **RNN** based forecasting strategy performed much better than other related methods.*

Bandit Algorithm	Forecasting Strategy	Average Gain%
EXP4.P	EXAMM-RNN	21.95
	GRU	20.59
	LSTM	18.71
	VAR	14.95
	ARIMA	17.77
	Trivial-RET	15.80
Thompson Sampling with Linear Payoff	EXAMM-RNN	4.67
	GRU	5.23
	LSTM	4.48
	VAR	4.33
	ARIMA	4.28
	Trivial-RET	4.42
LinUCB	EXAMM-RNN	0.70
	GRU	0.07
	LSTM	-0.55
	VAR	1.44
	ARIMA	-0.22
	Trivial-RET	0.65
Buy & Hold	-	16.78

Comparison of Bandit Algorithms with respect to average % gain earned across all 30 company stocks. Results for each forecasting method are present in the respective row. The Trivial-RET strategy covers a CMAB example, while the others strategies represent iCMABs.



1. **World Model** - The world model f_w is responsible for forecasting the context c_t itself given the previously encountered context c_{t-1} and encoding $E_t(.)$.

$$f_w(c_{t-1}, E_t(.)) = c_t$$

2. **Controller/Action Model** - The controller f_c is responsible for estimating the reward values μ_r for each action a_t, given the current context c_t and an encoding $E_t(.)$ of other signals that might help inform what action to take next.

$$f_c(a_t, c_t, E_t(.)) = \mu_r$$

Devroop Kar, Zimeng Lyu, Alexander G. Ororibia, Travis Desell, and Daniel Krutz. **Enabling An Informed Contextual Multi-Armed Bandit Framework For Stock Trading With Neuroevolution**. *Proceedings of the Genetic and Evolutionary Computation Conference Companion.* Melbourne, Australia. July 14-18, 2024.

Zimeng Lyu, Devroop Kar, Matthew Simoni, Rohaan Nadeem, Avinash Bhojanapalli, Hao Zhang and Travis Desell. **Evolving RNNs for Stock Forecasting: A Low Parameter Efficient Alternative to Transformers**. *The 28th International Conference on the Applications of Evolutionary Computation (EvoStar: EvoApps 2025)*. Trieste, Italy. April 23-25, 2025.

RIT

Award ID#: 2225354