Differentiable Rule Induction from Raw Sequence Inputs

Kun Gao[1], Katsumi Inoue[2], Yongzhi Cao[3], Hanpin Wang[3], Feng Yang[1]

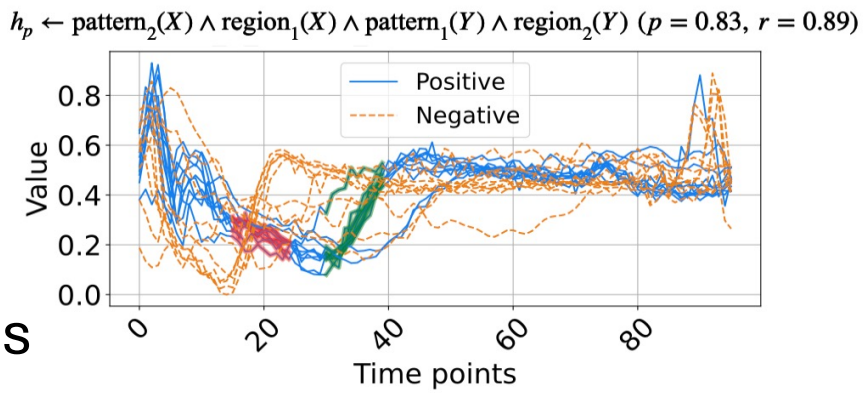
[1] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR)

[2] National Institute of Informatics

[3] School of Computer Science, Peking University

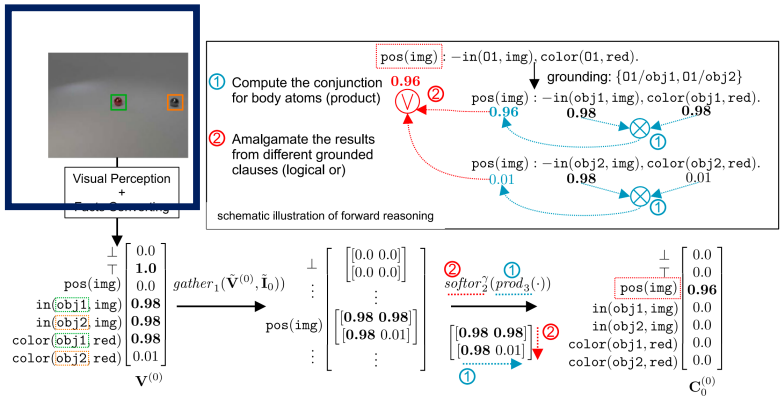
Motivation

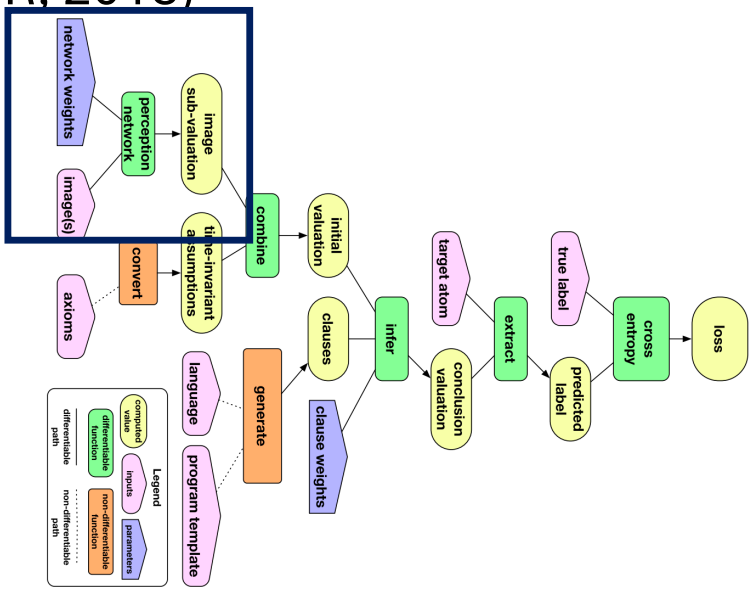
- Learning rules from raw sequence to explain the ground-truth class
- To explain the ground truth class with patterns
- The patterns are related to some actions
- This action can also be used to explain data
- The same idea can be applied to healthcare, finance, energy, etc.



Challenges

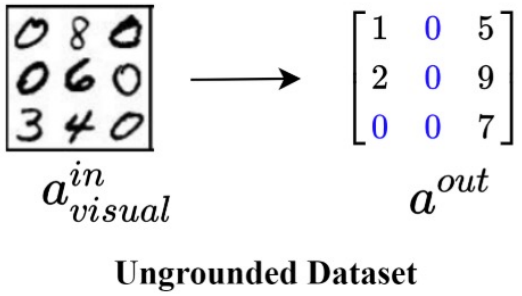
- Label leakage: Most neuro-symbolic methods use labels of components to implement inductive/deductive logic programming
  - αILP (Shindo et al., Machine Learning, 2023)
  - ∂ILP (Evans and Grefenstette, JAIR, 2018)



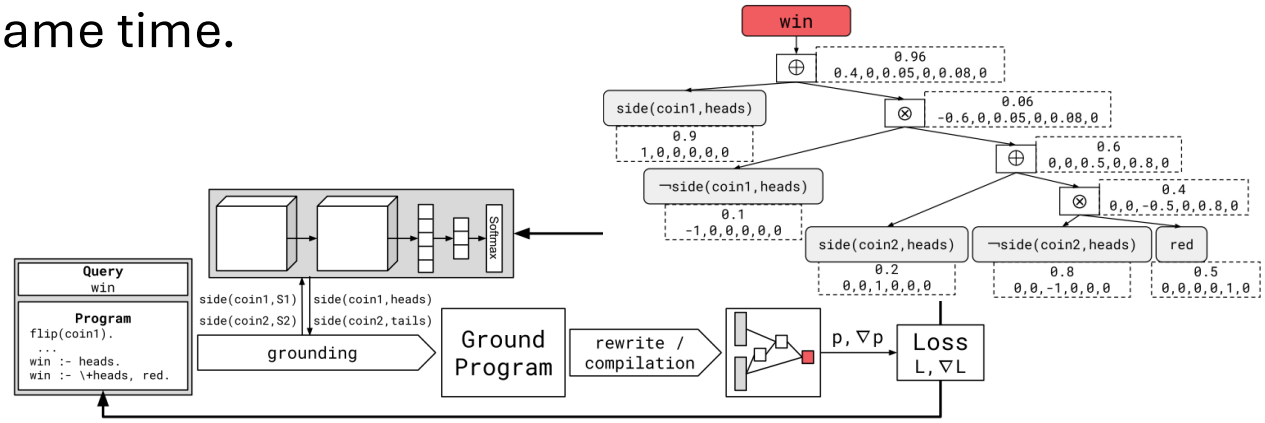


Related works

- SAT domain
  - SATNet (Topan et al.): Learning missing images of handwritten digits with neural networks.
- Logic programing domain
  - DeepProbLog (Manhaeve et al.): Training perception and the probabilities of target atom at the same time.



SATNet



(a) The learning pipeline.

DeepProbLog

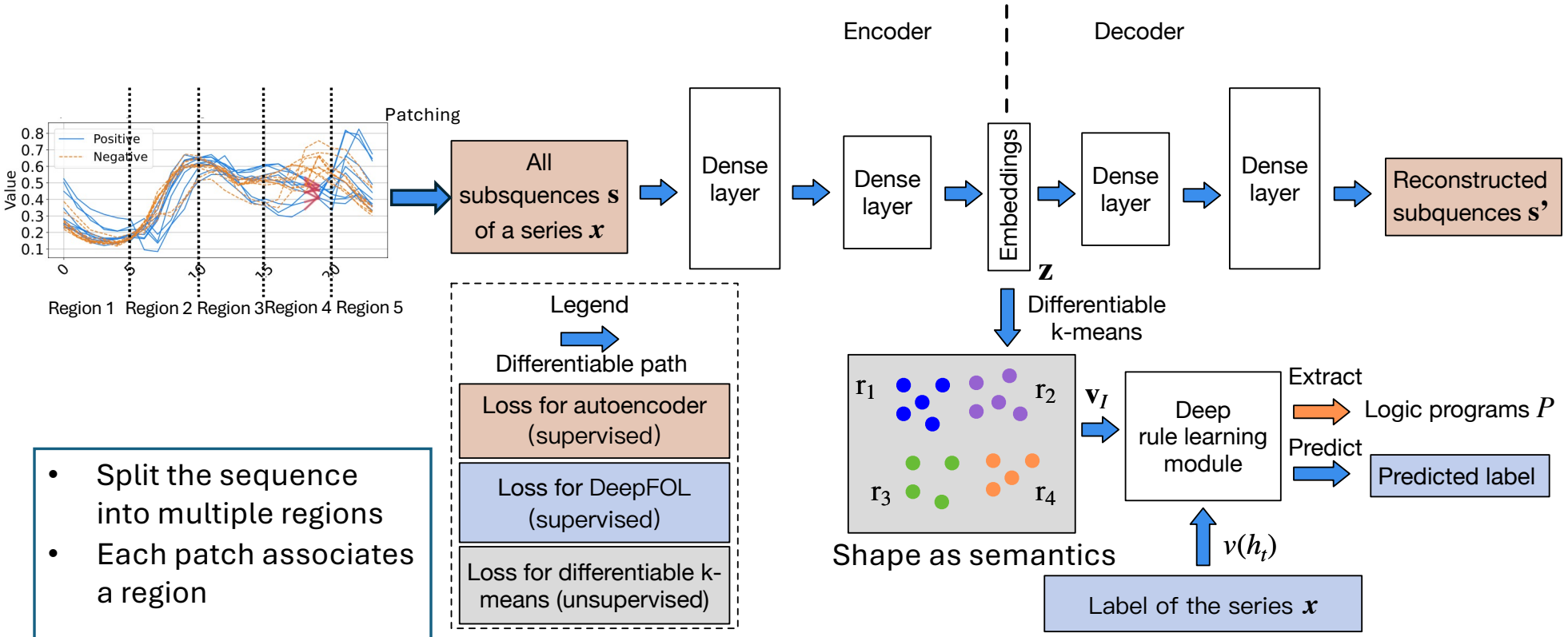
Language bias

- A variable X_j can be grounded with a feature
- Consider binary class, using h_t represent the positive class
- A ground-truth class sequence data can be described with its features:

$$h_t \leftarrow pattern_{i_1}(X_{j_1}), region_{k_1}(X_{j_1}), \ldots, pattern_{i_{n_1}}(X_{j_{n_2}}), region_{k_{n_3}}(X_{j_{n_2}}),$$

- The pattern predicate can express any discriminative shapes
- region predicates attached to pattern predicates can express
  - The temporal relations of two patterns is important
  - The specific time information of patterns is important
- One region corresponds to one pattern, and one pattern may occur in multiple regions

Methods: Clustering method and neural network



- Split the sequence into multiple regions
- Each patch associates a region

$$\min_{\mathcal{R},\gamma_e,\gamma_l} \sum_{\mathbf{s}\in\mathbf{x},\mathbf{x}\in\mathbf{X}} f_1\left(\mathbf{s}, A\left(\mathbf{s};\gamma_e\right)\right) + \lambda_1 \sum_{k=1}^{K} f_1\left(\mathbf{h}_{\gamma_e}(\mathbf{s}), \mathbf{r}_k\right) G_{k,f_1}\left(\mathbf{h}_{\gamma_e}(\mathbf{s}), \alpha; \mathcal{R}\right) + \lambda_2 f_2\left(N_R\left(\tilde{L}_b\left(\mathbf{x}\right);\gamma_l\right), y\right),$$

Clustering methods

- Differentiable clustering[1]:
  - K, number of clusters
  - γ, parameters in the autoencoder
  - f, distance function

$$\min_{\mathcal{R},\gamma} \sum_{\mathbf{x}\in\mathbf{X}} \underbrace{f(\mathbf{x}, A(\mathbf{x};\gamma))}_{\text{Autoencoder}} + \lambda \sum_{k=1}^{K} \underbrace{f(\underbrace{\mathbf{h}_\gamma(\mathbf{x})}_{\text{encoder}}, \underbrace{\mathbf{r}_k}_{\text{clusters}})}\underbrace{G_{k,f}(\mathbf{h}_\gamma(\mathbf{x}), \alpha; \mathcal{R})}_{\text{weights}},$$

- Differentiable weight function: Softmax function

$$G_{k,f}(\mathbf{h}_\gamma(\mathbf{x}), \alpha; \mathcal{R}) = \frac{e^{-\alpha f(\mathbf{h}_\gamma(\mathbf{x}), \mathbf{r}_k)}}{\sum_{k'=1}^{K} e^{-\alpha f(\mathbf{h}_\gamma(\mathbf{x}), \mathbf{r}_{k'})}},$$

- The sum of all weights is 1
- The smaller the distance, the larger the weight

[1] Maziar Moradi Fard, Thibaut Thonet, and Eric Gaussier. Deep k-means: Jointly clustering with k-means and learning representations. Pattern Recognit. Lett., 138:185–192, 2020.

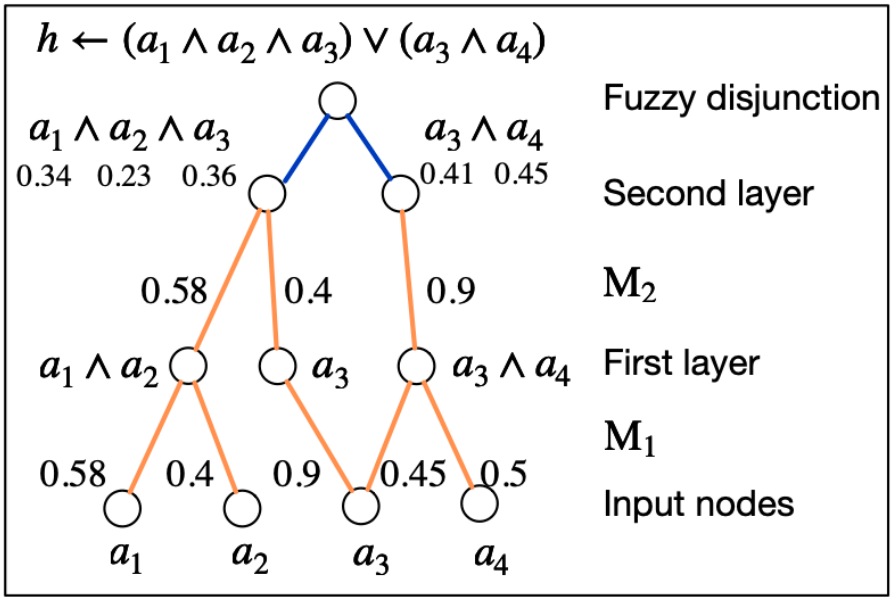
Rule learning module[1]

- The forward computation:

$$\hat{y} = \tilde{\bigvee}_{i=1}^{m} \left(g_k \circ g_{k-1} \circ \cdots \circ g_1(\mathbf{v}_I)\right)[i],$$

$$g_i(\mathbf{x}_{i-1}) = \frac{1}{1-d}\,\mathbf{ReLU}\left(\mathbf{M}_i\mathbf{x}_{i-1} - d\right),$$

$$\mathbf{M}_i[j,k] = \frac{e^{\tilde{\mathbf{M}}_i[j,k]}}{\sum_{k=1}^{n_{\text{in}}} e^{\tilde{\mathbf{M}}_i[a,k]}}.$$



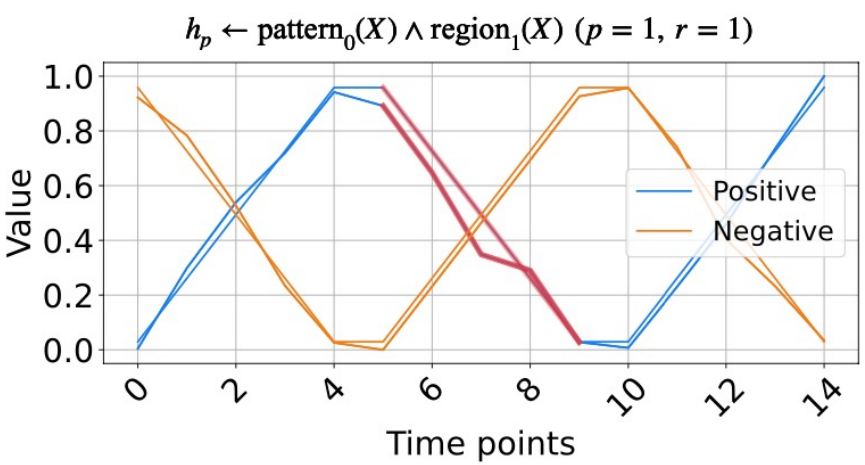
- Interpretable matrix $\mathbf{M}_P$
  - The product of each softmax-valued layer

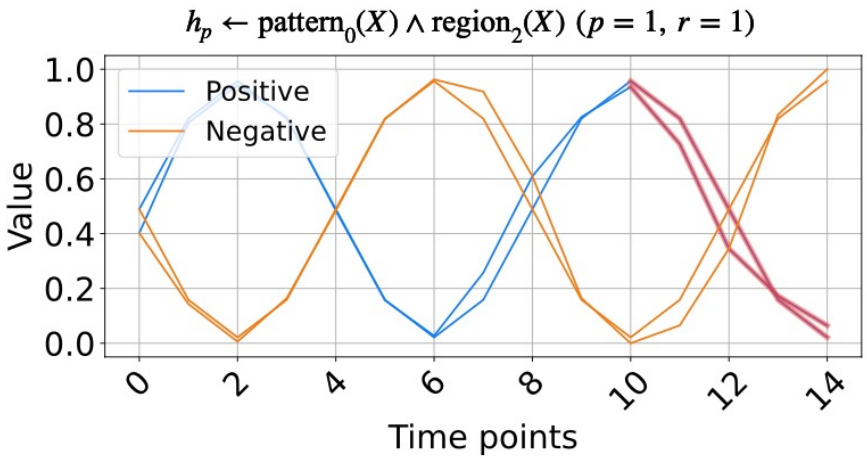
$$\mathbf{M}_P = \prod_{i=1}^{k} \mathbf{M}_i$$

[1] Kun Gao, Katsumi Inoue, Yongzhi Cao, and Hanpin Wang. A differentiable first-order rule learner for inductive logic programming. Artif. Intell., 331:104108, 2024.

Experiments: retrieve the pre-defined rules

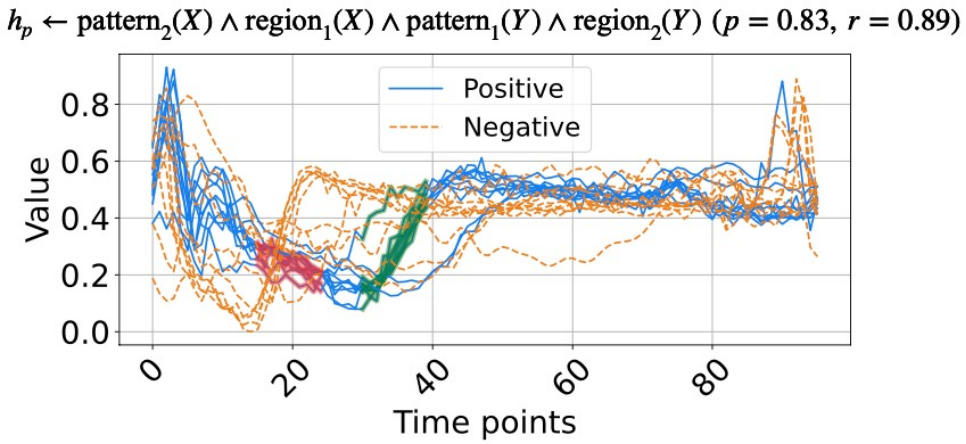
- Each class has two patterns: increasing and decreasing
- Each pattern has a length of 5
- The learned rules and the corrected patterns can be presented as follows[1]



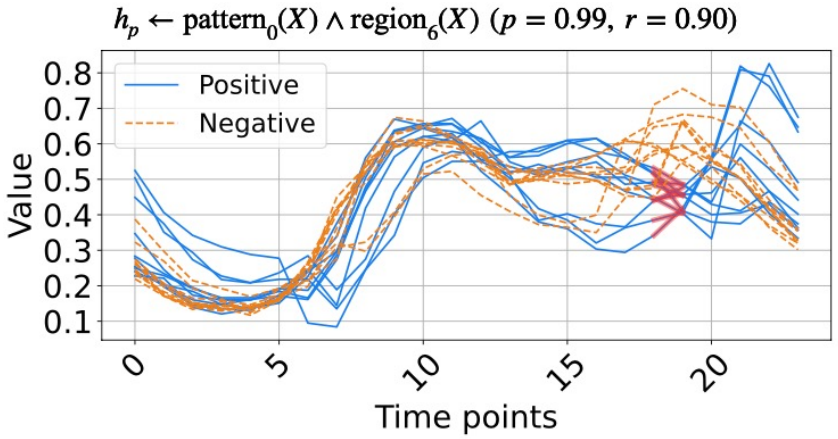


Experiments: learning from UCR archive

- We choose ECG and ItalyPow.Dem. from UCR binary variable time series data
- For ECG, a decrease (15 to 25) and then increase (30 to 40) is key information for positive class
- For ItalyPow.Dem., lower values around 18 to 19 is key information for positive class



(a) A rule from ECG dataset.



(b) A rule from ItalyPow.Dem. dataset.

Experiments: learning from UCR archive

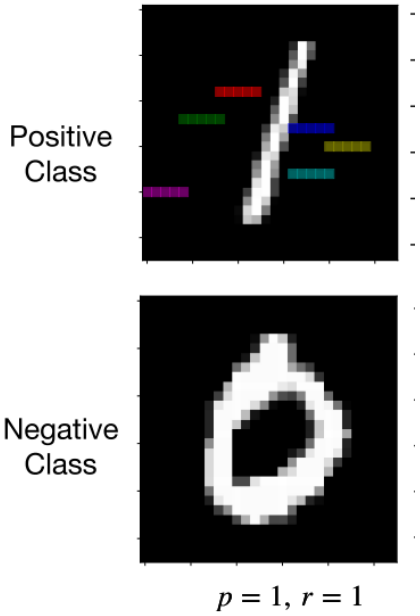
- Compare the classification accuracy of extracted rules (NeurRL(R)) from the model and neural networks (NeurRL(N)) with baselines
- The discrete rules decrease the performance of the model

Table 1: Classification accuracy on 13 binary UCR datasets with different models.

Dataset	C.	I.	Length	Xu	BoW	SSSL	NeurRL(R)	NeurRL(N)
Coffee	2	56	286	0.588	0.620	0.792	0.964	**1.000**
ECG	2	200	96	0.819	**0.955**	0.793	0.820	0.880
Gun point	2	200	150	0.729	**0.925**	0.824	0.760	0.873
ItalyPow.Dem.	2	1096	24	0.772	0.813	**0.941**	0.926	0.923
Lighting2	2	121	637	0.698	0.721	**0.813**	0.689	0.748
CBF	3	930	128	0.921	0.873	**1.000**	0.909	0.930
Face four	4	112	350	0.833	0.744	0.851	0.914	**0.964**
Lighting7	7	143	319	0.511	0.677	0.796	0.737	**0.878**
OSU leaf	6	442	427	0.642	0.685	0.835	0.844	**0.849**
Trace	4	200	275	0.788	**1.00**	**1.00**	0.833	0.905
WordsSyn	25	905	270	0.639	0.795	0.875	0.932	**0.946**
OliverOil	4	60	570	0.639	0.766	0.776	0.768	**0.866**
StarLightCurves	3	9236	2014	0.755	0.851	0.872	0.869	**0.907**
Mean accuracy				0.718	0.801	0.859	0.842	**0.891**

Experiments: Learning from images

- Transfer the image to sequence data
- Set the images with the number 1 as the positive class and the images with the number 0 as the negative class.
- Each highlighted sequence and its period (location) can distinguish the images with the number 1 from the images with the number 0.
- The precision = 1 and the recall = 1 for the learned rule presented on the side.



Experiments: learning from UCR archive

- Compare the accuracy and running time with non-differentiable k-means
- Differentiable k-means can incorporate the differentiable rule learning method to have a better performance based on the non-differentiable k-means with differentiable rule learning method

Table 2: Comparisons with non-differentiable k-means clustering algorithm.

Dataset	Non-differentiable k-means		Differentiable k-means	
	accuracy	running time (s)	accuracy	running time (s)
Coffee	0.893	313	**0.964**	**42**
ECG	0.810	224	**0.820**	**65**
Gun point	**0.807**	102	0.740	**35**
ItalyPow.Dem.	0.845	114	**0.926**	**63**
Lighting2	0.672	1166	**0.689**	**120**

Ablation study: Comparing accuracy with and without pre-training

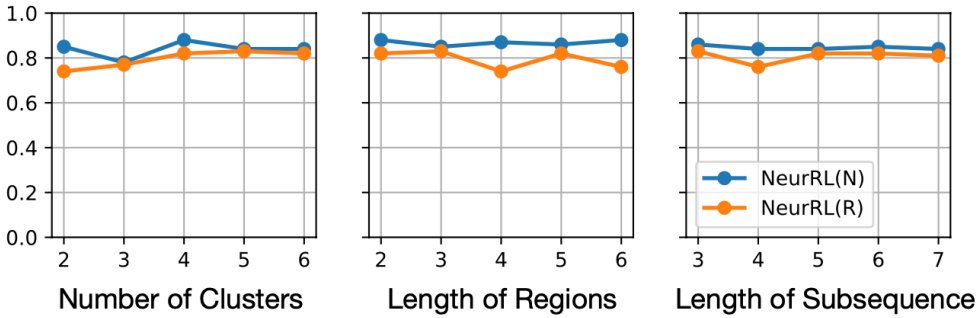
- With pre-training: Pre-train autoencoder and clustering method (k-means)
- Without pre-training: Train autoencoder, differentiable clustering method, and rule-learning module together

Dataset	With pre-training		Without pre-training	
	Accuracy	Running time (s)	Accuracy	Running time (s)
Coffee	**1.00, 0.96**	42	0.83, 0.81	**30**
ECG	**0.88, 0.82**	65	0.87, 0.64	**53**
ItalyPow.Dem.	**0.92, 0.93**	63	0.75, 0.80	**61**
Gun Point	**0.87, 0.76**	35	0.86, 0.43	**31**
Lighting2	**0.75, 0.69**	120	0.64, 0.60	**63**

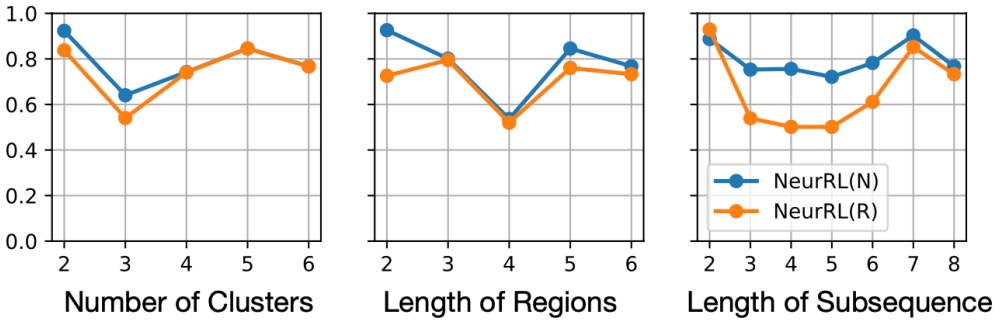
Table 3: Ablation study: With vs. without pre-training autoencoder and clustering model. The first accuracy is obtained by NeurRL(N) and the second accuracy is obtained by NeurRL(R).

Ablation study: Sensitivity of model

- The key hyperparameters
  - Number of clusters
  - Length of regions
  - Length of subsequences



(a) On ECG dataset.



(b) On ItalyPow.Dem. dataset.

Conclusion

Conclusion

- Learning from raw input without pre-defined object labels (avoid explicit supervised labels)
- Fully differentiable rule learning process
- Interpretability: Explain the data and model

Limitations

- Learning from time series data with missing values
- Learning from sequences of raw images

Thank you for your attention!