Deciphering Raw Data in Neuro-Symbolic Learning with Provable Guarantees

Lue Tao, Yu-Xuan Huang, Wang-Zhou Dai, Yuan Jiang

Shi-Jie Sang
April 18,2025

Conventional Supervised Learning

Let $\mathcal{X} \subset \mathbb{R}^d$ be the input space and $\mathcal{Y} = \{0, 1, \ldots, c-1\}$ be the label space. The objective is to learn a mapping $h : \mathcal{X} \to \mathbb{R}^c$ that minimises the expected risk:

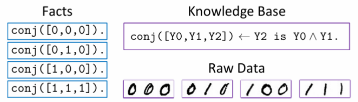
$$\mathcal{R}(h) = \mathbb{E}_{p(x,y)} \ell(h(x), y),$$

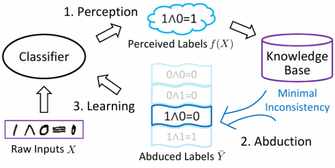
where $\ell : \mathbb{R}^c \times (y) \to \mathbb{R}$ is a loss function that measures how well the classifier perceives an input.

Neuro-Symbolic Learning

The raw inputs $X = [x_0, x_1, \ldots, x_{m-1}]$ are given, while their ground-truth $Y = [y_0, y_1, \ldots, y_{m-1}]$ are not observable. Instead, we only know that the logical facts grounded by the labels are compatible with a given knowledge base.

Example



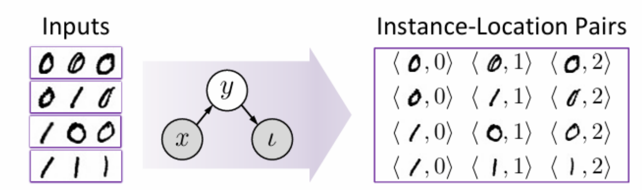


$$\mathcal{R}_{\mathsf{NeSy}}(h) = \mathbb{E}_{p(X,\tau)} \mathcal{L}(X, \bar{Y}; h), \mathsf{s.t.} B \cup \bar{Y} \models \tau$$

Heuristics to search for the most likely labels

- Maximal probability: $\bar{Y} = \mathrm{argmax}_Y \hat{p}(Y \mid X)$
- Minimal distance: $\bar{Y} = \mathrm{argmin}_Y ||Y - \hat{Y}||$
- Random: Randomly select an element Y from the candidate set
- ......

Instance-Location Pairs



From n sequences of unlabelled data $\{X^{(i)}\}_{i=0}^{n-1}$, a total of mn instance-location pairs $\{\langle x^{(i)}, \iota^{(i)}\rangle\}_{i=0}^{mn-1}$ will be obtained.

Location-Based Risk

Assumption. $p(\iota \mid y, x) = p(\iota \mid y)$.

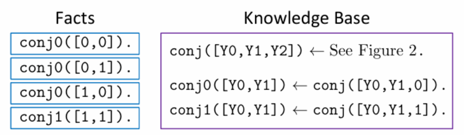
$$p(\iota = k \mid x) = \sum\nolimits_{j=0}^{c-1} Q_{jk} \cdot p(y = j \mid x),$$

where $Q_{jk} = p(\iota = k \mid y = j)$.

In practice, we let $q(x) = Q^{\mathrm{T}} g(x)$ and interpret $g(x)$ as probabilities via $g_j(x) = \exp(h_j(x)) / \sum_{i=0}^{c-1} \exp(h_i(x))$. Then we minimise the Location-based risk:

$$\mathcal{R}_{\mathrm{L}}(h) = \mathbb{E}_{p(x,\iota)} \ell(q(x), \iota).$$

Example



A set of mn triplets $\{\langle x^{(i)}, \iota^{(i)}, \tau^{(i)} \rangle\}_{i=0}^{mn-1}$ will be obtained. For conciseness, we use $\widetilde{y}$ to denote $\langle \iota, \tau \rangle$. Then, the triplets above becomes $\{\langle x^{(i)}, \widetilde{y}^{(i)} \rangle\}_{i=0}^{mn-1}$. Similar to before, we have

$$p(\widetilde{y} = o \mid x) = \sum_{j=0}^{c-1} \widetilde{Q}_{jo} \cdot p(y = j \mid x),$$

where $\widetilde{Q}_{jo} = p(\widetilde{y} = o \mid y = j)$ and $o = tm + k$. Then we'll have target-location-based risk:

$$\mathcal{R}_{\mathrm{TL}}(h) = \mathbb{E}_{p(x,\widetilde{y})} \ell(\widetilde{q}(x), \widetilde{y}).$$

Theorem

Assumption. $\forall Y \in \mathcal{S}(\tau), p(Y) = 1/|\mathcal{S}(\tau)|$.

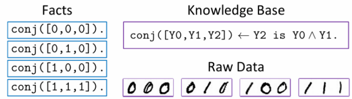
Theorem 1. $\mathcal{R}_{\mathrm{L}}(h) \leq \mathcal{R}_{\mathrm{NeSy}}(h) + C(\mathcal{S}(\tau))$.

Theorem 2. $\mathcal{R}_{\mathrm{TL}}(h) \leq \mathcal{R}_{\mathrm{NeSy}}(h) + C(p(\tau))$.

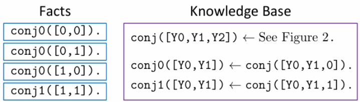
Theorem 3. If $\widetilde{Q}$ has full row rank, then $h^*_{\mathrm{TL}} = \mathrm{argmin}_h \mathcal{R}_{\mathrm{TL}}(h)$ recovers $h^* = \mathrm{argmin}_h \mathcal{R}(h)$, *i.e.*, $h^*_{\mathrm{TL}} = h^*$.

Corollary 1. $h^*_{\mathrm{L}} = h^*$.

Example



$$Q = \begin{pmatrix} 2/7 & 2/7 & 3/7 \\ 2/5 & 2/5 & 1/5 \end{pmatrix}$$



$$Q = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

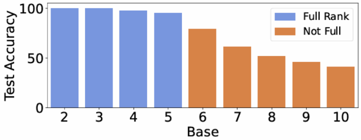
$$\widetilde{Q} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

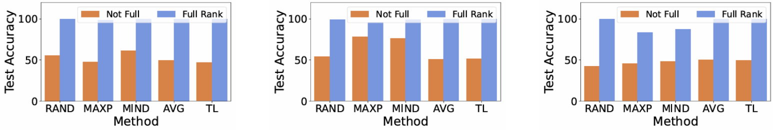
Experiment

Task	Method	MNIST	EMNIST	USPS	Kuzushiji	Fashion
ConjEq	Rand	99.91 ± 0.06	99.65 ± 0.04	99.33 ± 0.16	97.82 ± 0.35	98.40 ± 0.11
	MaxP	99.94 ± 0.04	99.82 ± 0.03	99.20 ± 0.00	98.80 ± 0.16	99.39 ± 0.12
	MinD	99.91 ± 0.08	99.84 ± 0.07	99.14 ± 0.17	98.91 ± 0.17	98.84 ± 0.19
	Avg	99.85 ± 0.10	99.80 ± 0.07	99.30 ± 0.17	98.34 ± 0.16	98.62 ± 0.21
	TL	99.92 ± 0.05	99.82 ± 0.06	99.25 ± 0.08	98.53 ± 0.26	98.77 ± 0.06
Conjunction	Rand	99.91 ± 0.06	99.86 ± 0.04	99.30 ± 0.13	98.79 ± 0.13	99.00 ± 0.27
	MaxP	99.93 ± 0.04	99.81 ± 0.02	99.20 ± 0.00	98.62 ± 0.15	99.05 ± 0.09
	MinD	99.94 ± 0.02	99.79 ± 0.02	99.20 ± 0.00	98.74 ± 0.10	99.08 ± 0.10
	Avg	99.94 ± 0.02	99.85 ± 0.03	99.30 ± 0.13	98.68 ± 0.33	99.23 ± 0.13
	TL	99.94 ± 0.02	99.83 ± 0.04	99.20 ± 0.00	98.87 ± 0.18	99.30 ± 0.04
Addition	Rand	92.01 ± 0.93	92.94 ± 1.45	90.96 ± 1.04	73.18 ± 0.71	79.08 ± 2.61
	MaxP	96.40 ± 4.04	95.09 ± 5.20	94.29 ± 0.27	90.00 ± 0.27	87.34 ± 2.93
	MinD	98.32 ± 0.04	98.61 ± 0.06	94.61 ± 0.17	90.85 ± 0.26	88.40 ± 0.62
	Avg	94.90 ± 0.39	95.71 ± 0.42	93.22 ± 0.30	80.94 ± 0.62	84.43 ± 0.92
	TL	98.00 ± 0.14	98.41 ± 0.05	94.68 ± 0.20	90.04 ± 0.32	88.38 ± 0.25
HED	Rand	99.89 ± 0.02	99.71 ± 0.12	99.25 ± 0.23	97.68 ± 0.70	98.43 ± 0.55
	MaxP	99.90 ± 0.02	99.77 ± 0.02	99.23 ± 0.05	98.55 ± 0.08	99.33 ± 0.10
	MinD	99.87 ± 0.07	99.77 ± 0.02	99.21 ± 0.00	98.61 ± 0.21	99.32 ± 0.10
	Avg	99.60 ± 0.09	99.38 ± 0.21	99.32 ± 0.14	96.16 ± 1.22	98.46 ± 0.33
	TL	99.90 ± 0.02	99.77 ± 0.04	99.21 ± 0.00	98.50 ± 0.16	99.21 ± 0.06

Table 1: Test accuracy (%) of each method using MLP on benchmark datasets and tasks.

Experiment





(a) DNF, $m = 3$ (b) DNF, $m = 4$ (c) DNF, $m = 5$

My Works

What if there are less columns than rows in Q?

$$\mathrm{SimilarityScore}(h) = \mathrm{Inter}(h) - \mathrm{Intra}(h)$$

$$\mathrm{Inter}(h) = \sum_{i_1 \neq i_2, j_1 \neq j_2} \omega_{i_1 j_1} \omega_{i_2 j_2} \mathrm{Dis}(x_{i_1}, x_{i_2})$$

$$\mathrm{Inter}(h) = \sum_{i_1 \neq i_2, j} \omega_{i_1 j} \omega_{i_2 j} \mathrm{Dis}(x_{i_1}, x_{i_2})$$

If any two rows of Q are linearly independent. By maximising the SimilarityScore after minimising $\mathcal{R}_{\mathrm{L}}$, we may still obtain a good classifier.

Thm 0. 令 SimilarityScore 最小等价于令 $Ss = \sum_{i_1, i_2} \mathrm{Dis}(x_{i_1}, x_{i_2}) \cdot \omega_{i_1}\omega_{i_2}$ 最小.

Thm 1. 若距离的定义足够好，即任意相同类图像间距离均小于不同类图像间距离，并且训练集中，属于每一类的图像一样多时，那么当 h 将不同类图像完全区分开时，Ss 取最小值。

Thm 2. 令 $\omega = \widetilde{\omega}$ 时，Ss 可被 Ss 的最小值控制。

Thank you

Thank you!