NEURAL INFORMATION PROCESSING SYSTEMS

Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022)

University of Science and Technology of China (USTC)

DARE: Disentanglement-Augmented Rationale Extraction

Linan Yue[1,2], **Qi Liu**[1,2*], **Yichao Du**[1,2], **Yanqing An**[1,2], **Li Wang**[1,2,3], **Enhong Chen**[1,2]

1: Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China

2: State Key Laboratory of Cognitive Intelligence

3: ByteDance

{lnyue,duyichao,anyq,wl063}@mail.ustc.edu.cn;
{qiliuql,cheneh}@ustc.edu.cn

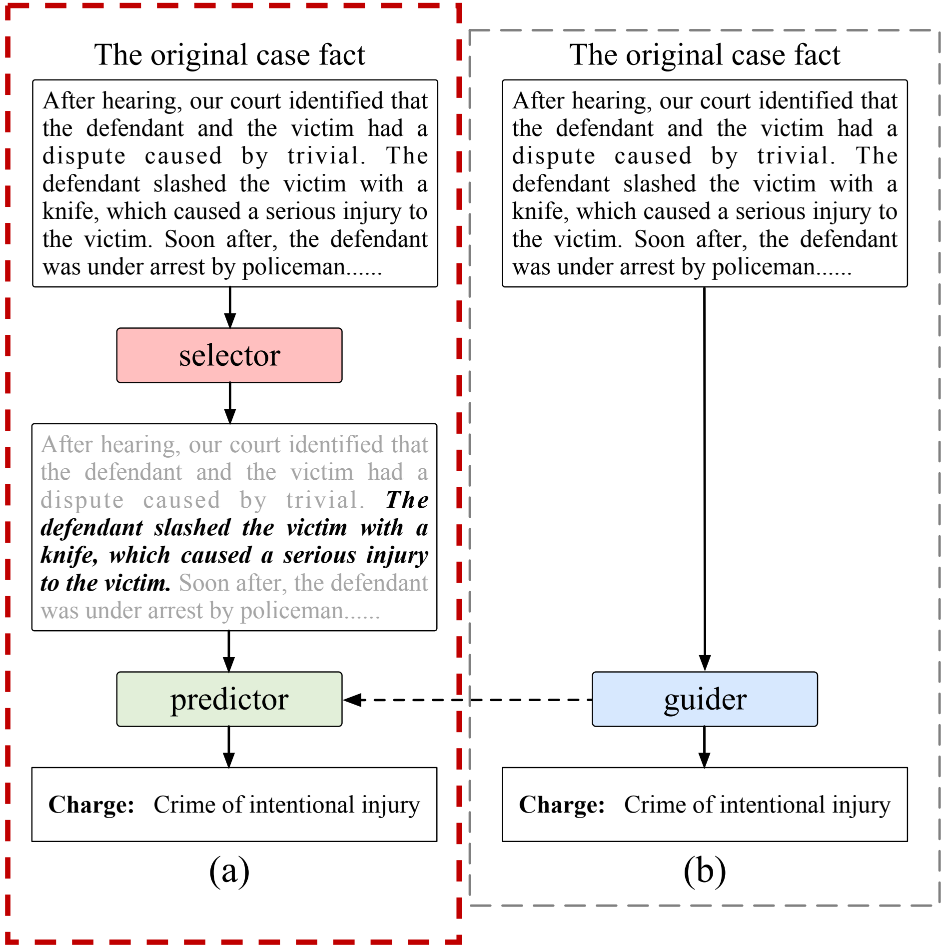
Presented by : Linan Yue

Background

Rationale Extraction

➢ Rationale Extraction

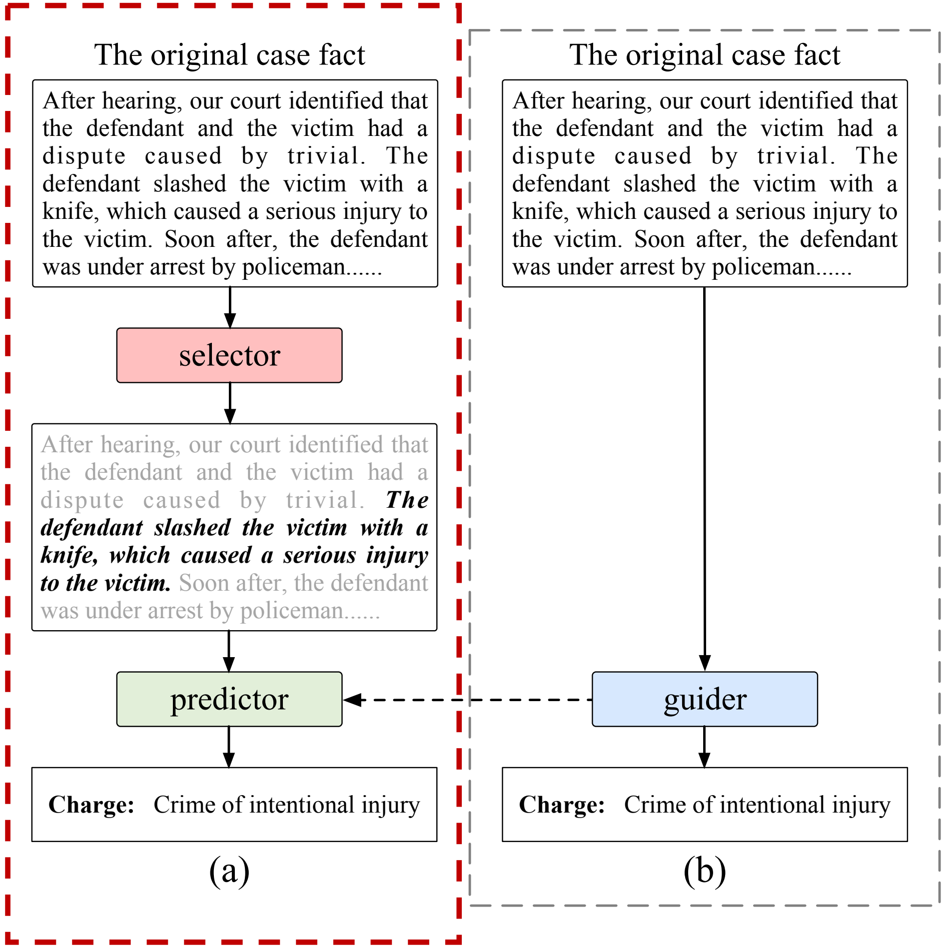
- It extracts a short and coherent part of original inputs (i.e., ***rationale***) as an explanation to support the prediction results when yielding them.



Background

Rationale Extraction

- **Rationale Extraction**
  - It extracts a short and coherent part of original inputs (i.e., ***rationale***) as an explanation to support the prediction results when yielding them.
- **Types of Rationale Extraction**
  - Traditional rationale extraction approaches cascade the ***selector*** and the ***predictor***.



Background

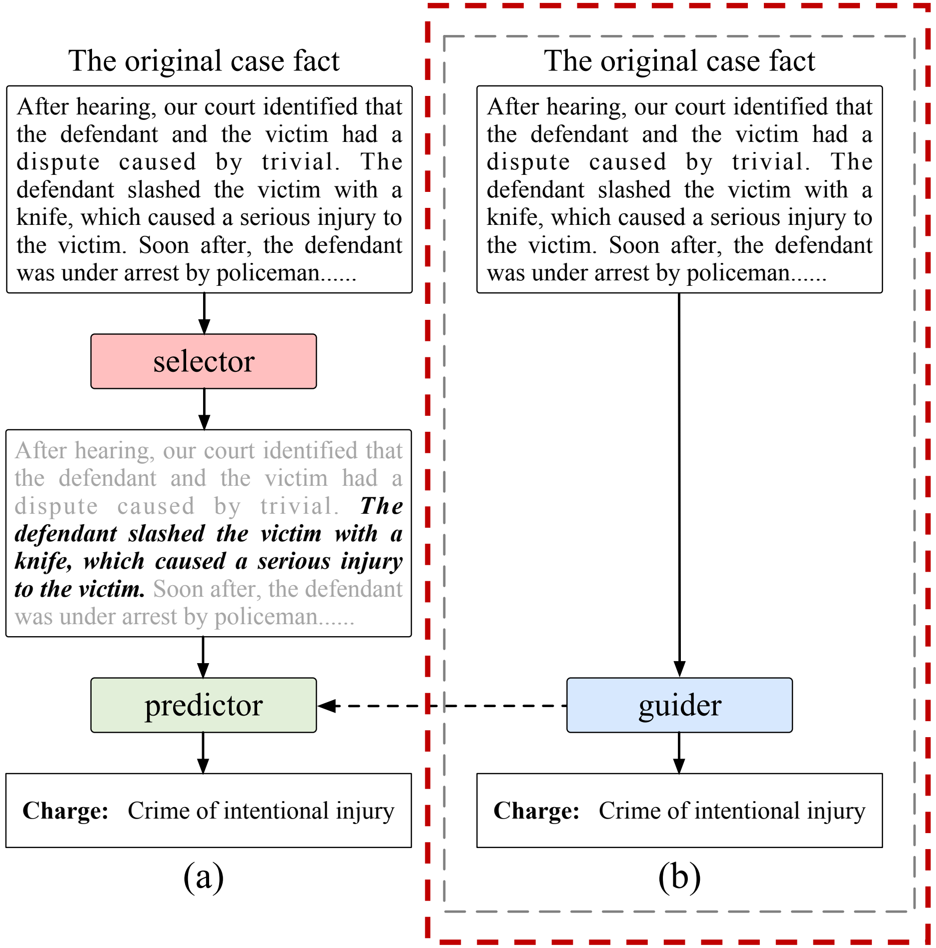
Rationale Extraction

➢ Rationale Extraction

- It extracts a short and coherent part of original inputs (i.e., ***rationale***) as an explanation to support the prediction results when yielding them.

➢ Types of Rationale Extraction

- Traditional rationale extraction approaches cascade the *selector* and the *predictor*.
- guidance pattern: adding an external ***guider***

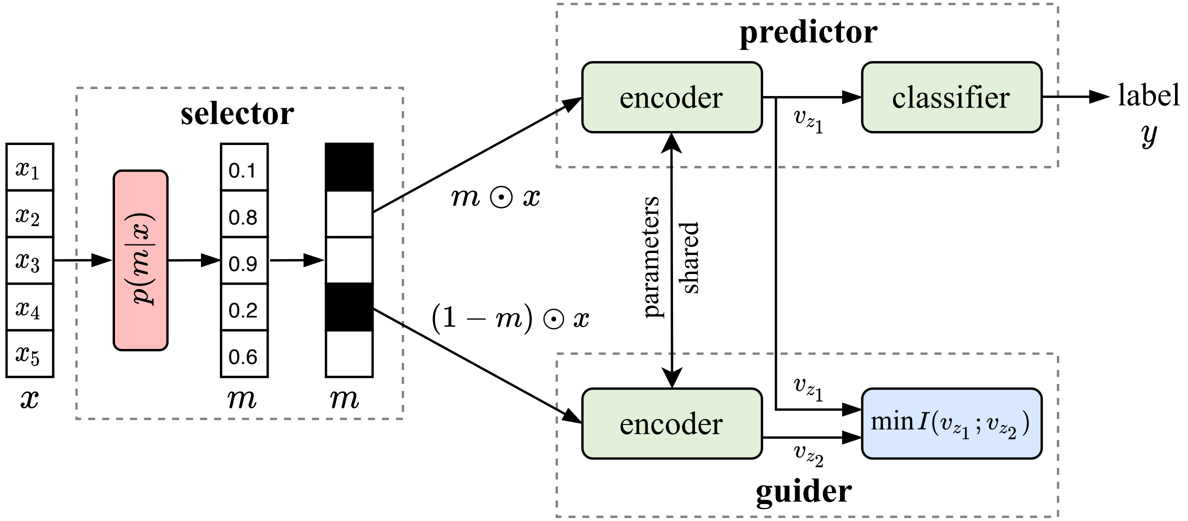


DARE

DARE

➢ Architecture of DARE

- ***self-guided***: Different from the previous model that requires external guidance, DARE aims to guide itself to extract more comprehensive rationales by squeezing more information from the input.

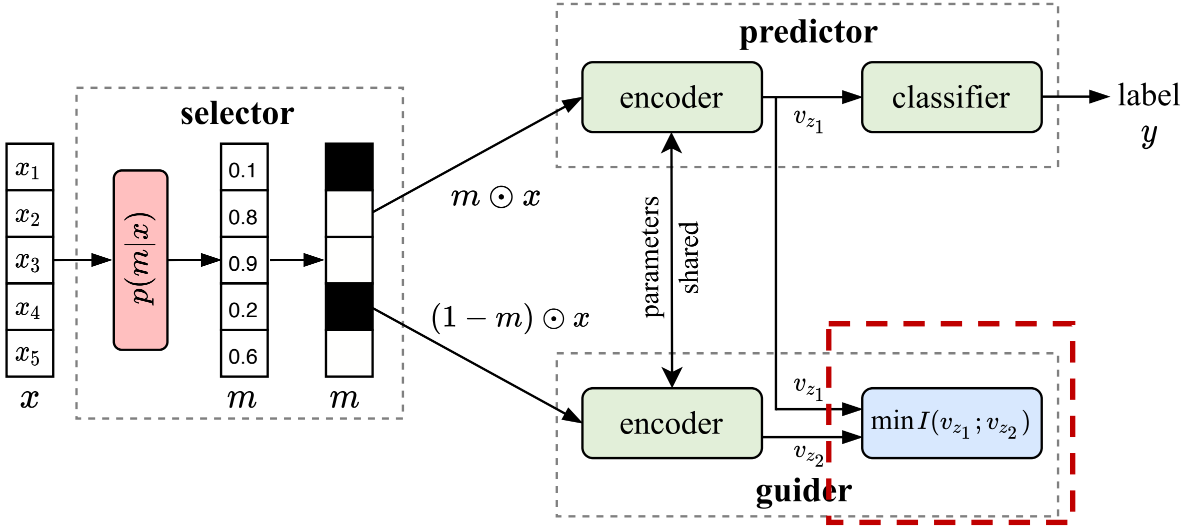


DARE

DARE

Architecture of DARE

- *self-guided*: Different from the previous model that requires external guidance, DARE aims to guide itself to extract more comprehensive rationales by squeezing more information from the input.
- Disentangled representations learning with ***mutual information*** minimization:



MI minimization

Mutual Information Estimation

- Mutual Information:

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right]$$

- InfoNCE: MI maximization

$$I_{nce} = \frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{f(x_i,y_i)}}{\frac{1}{N}\sum_{j=1}^{N} e^{f(x_i,y_j)}} = \frac{1}{N}\sum_{i=1}^{N} f(x_i,y_i) - \frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{1}{N}\sum_{j=1}^{N} e^{f(x_i,y_j)}\right]$$

- CLUB: MI minimization

$$I_{club} = \frac{1}{N}\sum_{i=1}^{N}\log p(y_i|x_i) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log p(y_j|x_i)$$

DARE

MI minimization

➢ CLUB_NCE

$$I_{nce} = \frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{f(x_i,y_i)}}{\frac{1}{N}\sum_{j=1}^{N}e^{f(x_i,y_j)}} = \frac{1}{N}\sum_{i=1}^{N}f\left(x_i,y_i\right) - \frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{1}{N}\sum_{j=1}^{N}e^{f(x_i,y_j)}\right]$$

$$I_{club} = \frac{1}{N}\sum_{i=1}^{N}\log p(y_i|x_i) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log p(y_j|x_i)$$

- Applying the Jensen's inequality ：

$$I_{nce} \leq \frac{1}{N}\sum_{i=1}^{N}f\left(x_i,y_i\right) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\left[e^{f(x_i,y_j)}\right] = \frac{1}{N}\sum_{i=1}^{N}f\left(x_i,y_i\right) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}f\left(x_i,y_j\right)$$

Experiments

Rationale evaluation

- Are the rationales extracted plausible and comprehensive ?
- Is the disentanglement operation effective ?
- Is CLUB_NCE effective on estimating mutual information ?

Experiments

Rationale evaluation

- **Are the rationales extracted plausible and comprehensive ?**

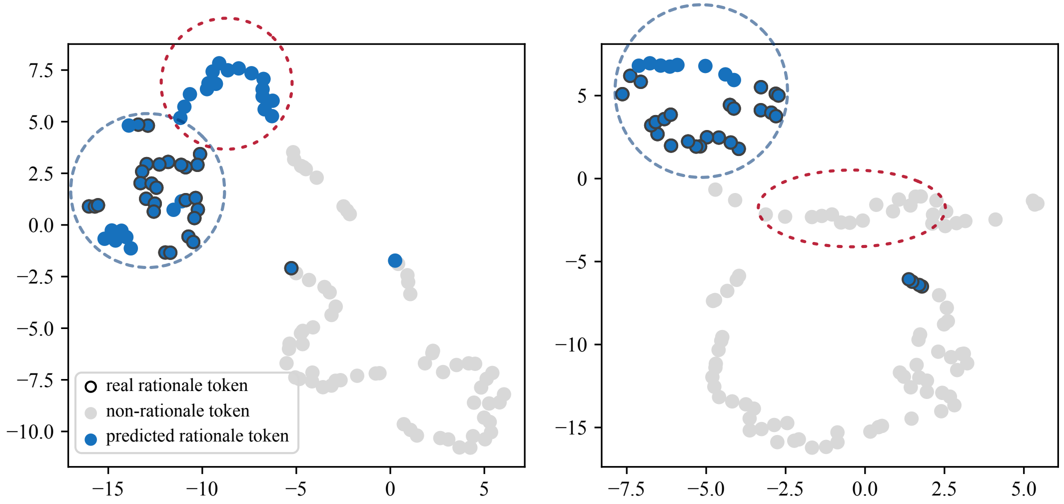
Table 1: Precision, Recall and F1 of selected rationales for three aspects. Among them, "% selected" represents the average proportion of selected tokens in the original text.

Methods	Appearance				Smell				Palate			
	Precision	Recall	F1	% selected	Precision	Recall	F1	% selected	Precision	Recall	F1	% selected
Bernoulli	96.3	56.5	71.2	14	95.1	38.2	54.5	7	80.2	53.6	64.3	7
HardKuma	**98.1**	65.1	78.3	13	**96.8**	31.5	47.5	7	**89.8**	48.6	63.1	7
InfoCal_IB	97.3	67.8	79.9	13	94.3	34.5	50.5	7	89.6	51.2	65.2	7
InfoCal(HK)	97.9	71.7	82.8	13	94.8	42.3	58.5	7	89.4	56.9	69.5	7
DARE (L1Out)	91.5	26.7	41.3	13	84.0	38.0	52.3	7	55.4	57.0	56.2	7
DARE (CLUB)	93.7	73.0	82.1	13	90.9	42.9	58.3	7	88.7	54.3	67.4	7
DARE	95.1	**73.5**	**82.9**	13	88.6	**46.8**	61.2	7	85.6	**59.0**	**69.9**	7
(std)	±0.2	±0.3	±0.1	-	±0.8	±0.6	±0.6	-	±0.6	±0.5	±0.2	-

Experiments

Rationale evaluation

➢ **Is the disentanglement operation effective ?**



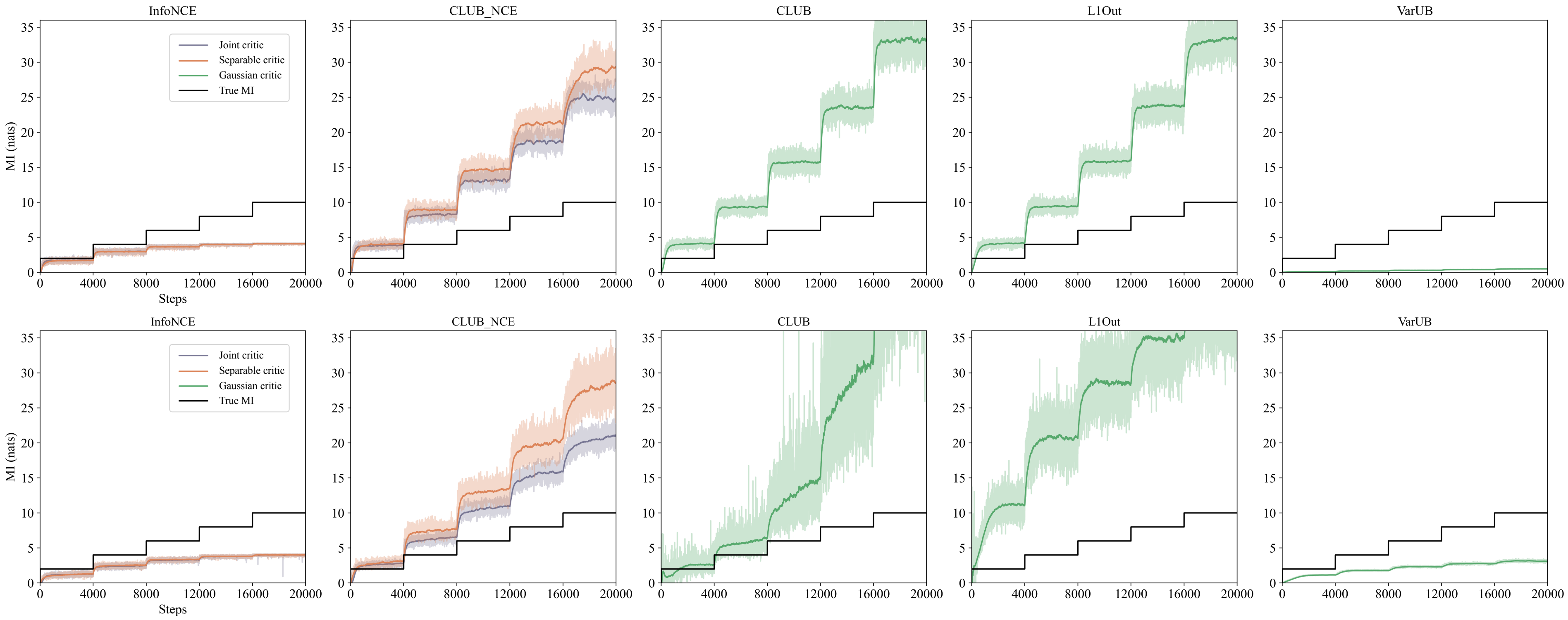
(a) HardKuma (b) DARE

Euclidean(**EU**) distance: **7.62** **12.18**

Experiments

MI evaluation

➢ **Is CLUB_NCE effective on estimating mutual information ?**



Thank you for listening!