The Twelfth International Conference on Learning Representations

认知智能全国重点实验室
STATE KEY LABORATORY OF COGNITIVE INTELLIGENCE

Towards Faithful Explanations: Boosting Rationalization with Shortcuts Discovery

Linan Yue[1], Qi Liu[1,2*], Yichao Du[1], Li Wang[3], Weibo Gao[1], Yanqing An[1]
1: State Key Laboratory of Cognitive Intelligence,
University of Science and Technology of China
2: Institute of Artificial Intelligence, Hefei Comprehensive National Science Center Hefei, China
3: ByteDance
{lnyue,duyichao,wl063,weibogao,anyq}@mail.ustc.edu.cn;
qiliuql@ustc.edu.cn

Presented by : Linan Yue

Outline

1 Background of Selective Rationalization

2 Shortcuts-fused Selective Rationalization

3 Experiments of SSR

Background

Selective Rationalization

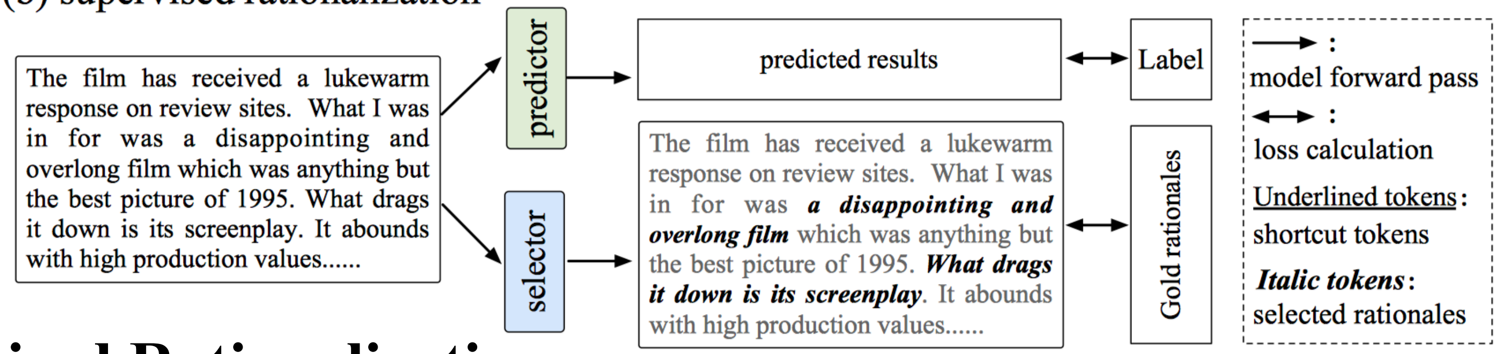
➢ Unsupervised Rationalization

- This type trains the selector and predictor in tandem.
- It is worth noting that the gold rationale is unavailable during the whole training process.

➢ Supervised Rationalization

- It models the rationalization with a multi-task learning, optimizing the joint likelihood of class labels and extractive rationales.

(b) supervised rationalization



➢ Semi-Supervised Rationalization

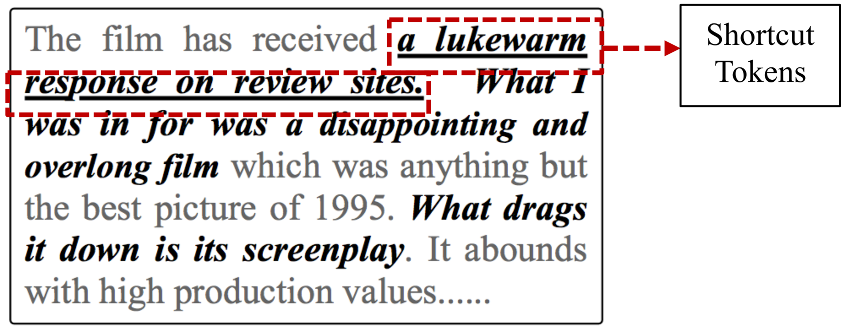
- Combining the superiority of the above two types of methods.

Background

Existing Problems

➢ **Low Faithfulness**

It is easy to exploit spurious correlations (aka., shortcuts) to yield the prediction results and compose the rationales.

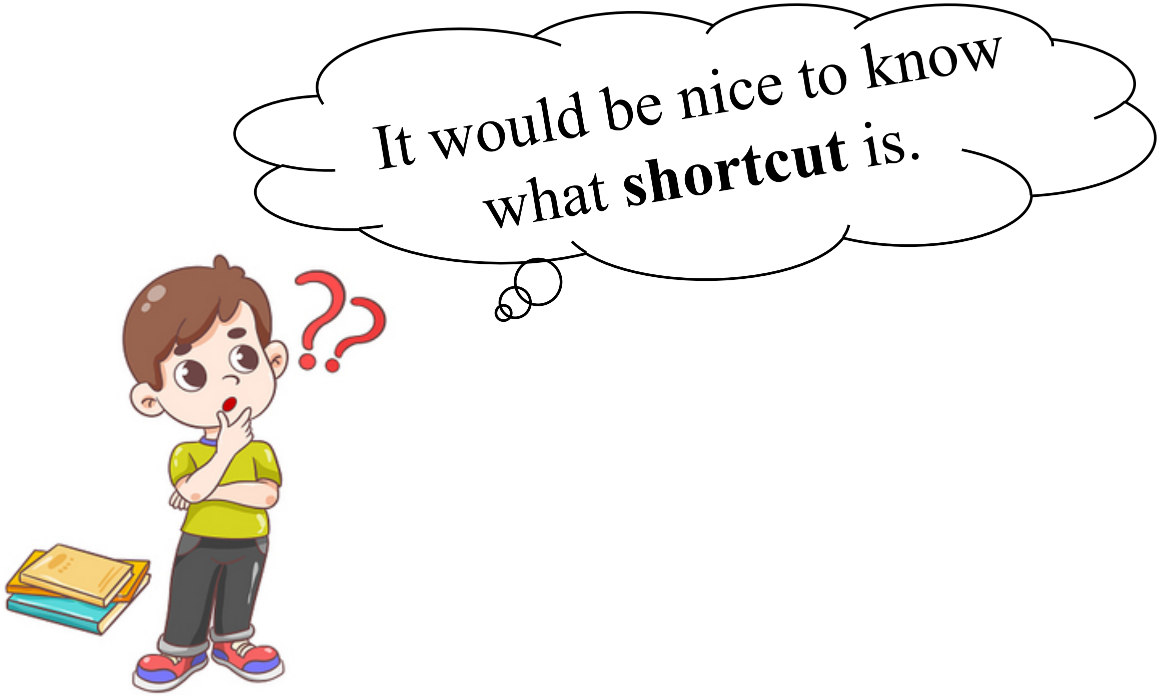


- Exploiting real rationales, the problem of adopting the shortcuts to predict task results can be mitigated.
- However, such extensive annotated rationales are infeasible to obtain for most tasks, rendering this method unavailable.

Background

Existing Problems

➢ **How to solve these problems?**



Outline

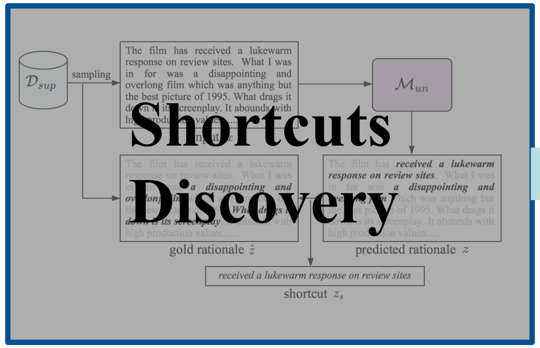
Shortcuts-fused Selective Rationalization

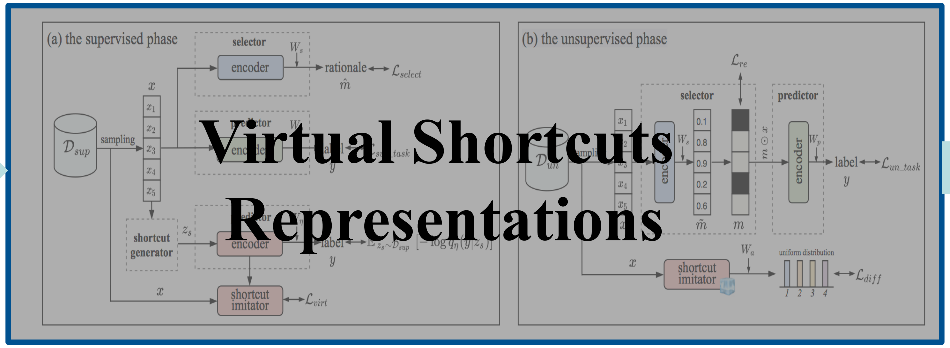
SSR

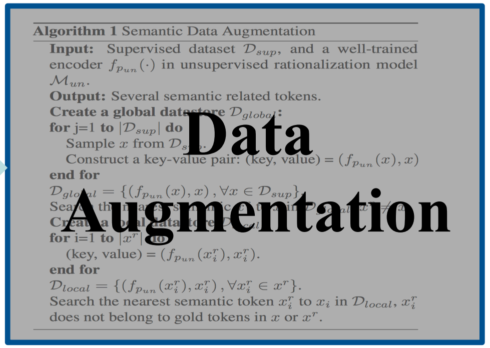
- **Semi-supervised rationalization**

  Considering a low-resource setup where they have annotated rationales for part of the training data $\mathcal{D}_{semi}$, $\mathcal{D}_{semi}$ consists of $\mathcal{D}_{un}$ and $\mathcal{D}_{sup}$, where $|\mathcal{D}_{un}| \gg |\mathcal{D}_{sup}|$.

Identifing potential **shortcuts**







Shortcuts-fused Selective Rationalization

SSR

➢ Shortcuts Discovery

- How to identify potential shortcuts?

Assumption 1: *A well-trained unsupervised rationalization model inevitably composes rationales with both the gold rationale and shortcuts tokens.*

Definition 1 *(Potential Shortcut Token) We first assume the unsupervised rationalization model $\mathcal{M}_{un}$ is already trained. Then, given the annotated rationales $\hat{z}$ and rationales z extracted by $\mathcal{M}_{un}$, we define $\mathbb{PST}(x_i)$ as whether a token x_i is considered to be a potential shortcut token or not:*

$$\mathbb{PST}(x_i) = \mathbb{I}(x_i \in z \wedge x_i \notin \hat{z}), \tag{5}$$

where $\wedge$ is the logical operation AND. $\mathbb{PST}(x_i)$=1 denotes x_i is defined as a potential shortcut token.

Shortcuts-fused Selective Rationalization

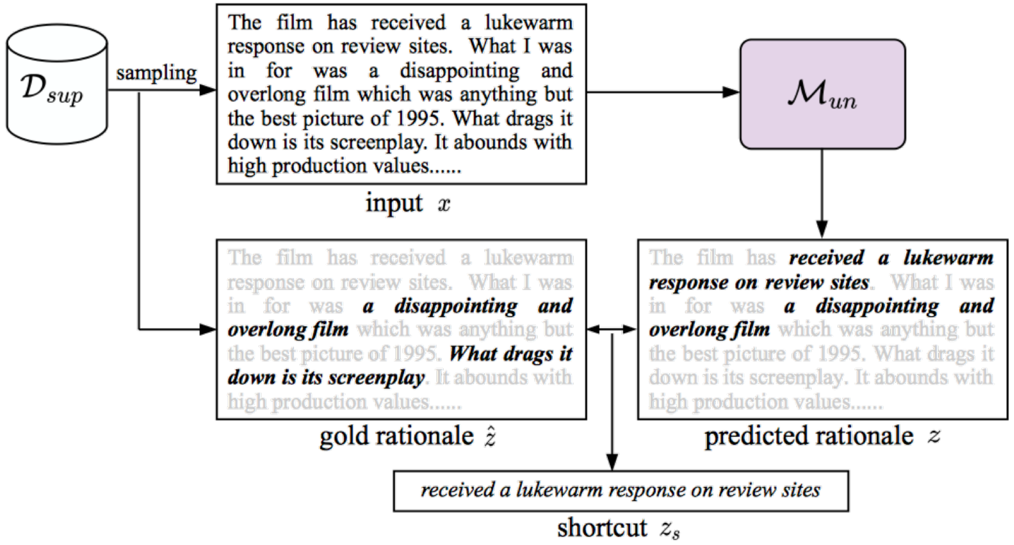
SSR

➢ Shortcuts Discovery

Definition 1 *(Potential Shortcut Token) We first assume the unsupervised rationalization model $\mathcal{M}_{un}$ is already trained. Then, given the annotated rationales $\hat{z}$ and rationales z extracted by $\mathcal{M}_{un}$, we define $\mathbb{PST}(x_i)$ as whether a token x_i is considered to be a potential shortcut token or not:*

$$\mathbb{PST}(x_i) = \mathbb{I}(x_i \in z \wedge x_i \notin \hat{z}), \tag{5}$$

where $\wedge$ is the logical operation AND. $\mathbb{PST}(x_i)$=1 denotes x_i is defined as a potential shortcut token.

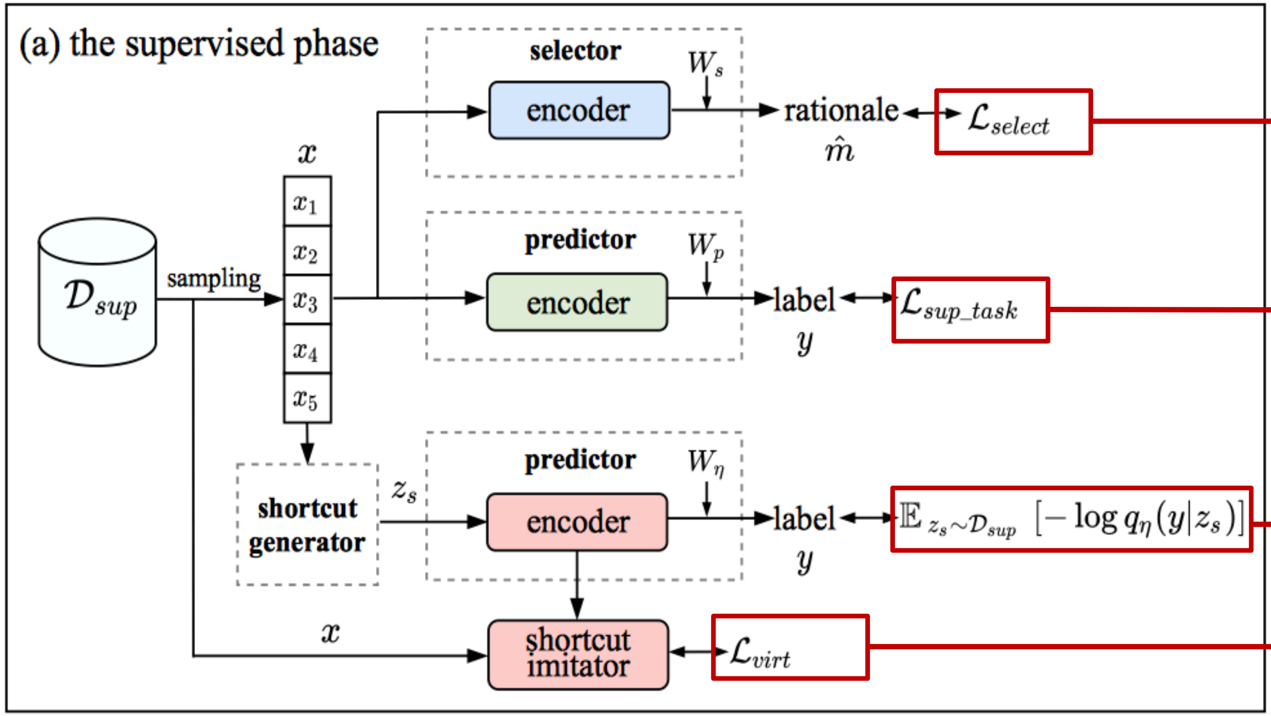


Shortcuts-fused Selective Rationalization

SSR

- **Virtual Shortcuts Representations**

In the supervised phase:



Token-level binary cross-entropy

$$\mathcal{L}_{select} = \sum_{i=1}^{n} -\hat{m}_i \log p_\theta(\widetilde{m}_i|x_i)$$

Task-level binary cross-entropy

$$\mathcal{L}_{sup\_task} = \mathbb{E}_{x,y\sim\mathcal{D}_{sup}}\left[-\log q_\psi(y|x)\right]$$

Capture sufficient shortcuts representations

$$\min \mathbb{E}_{z_s\sim\mathcal{D}_{sup}}\left[-\log q_\eta(y|z_s)\right]$$

Align and mimic shortcuts representations

$$\mathcal{L}_{virt} = \mathbb{E}_{x_{sup},z_s\sim\mathcal{D}_{sup}}\left[\left\| f_{p_\eta}(z_s) - f_a(x_{sup})\right\|^2\right]$$

Shortcuts-fused Selective Rationalization

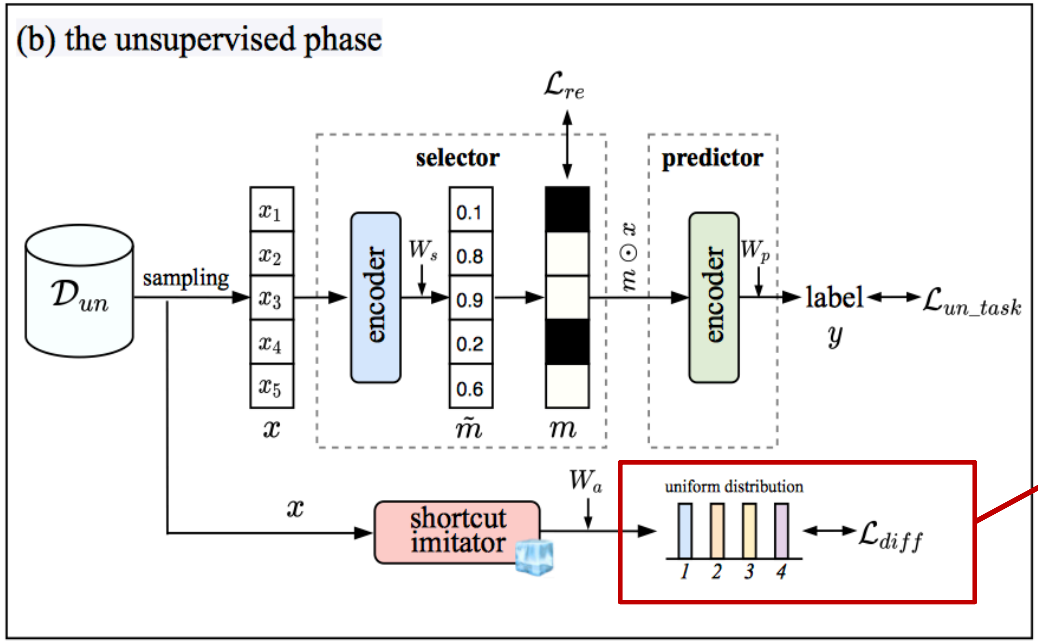
SSR

- **Virtual Shortcuts Representations**

  In the unsupervised phase:

  Generate virtual shortcuts representations

  Encourage the model to remove the effect of shortcuts on task predictions



$$\mathcal{L}_{diff} = \mathbb{E}_{x \sim \mathcal{D}_{un}} \left[\mathrm{KL}\left(\mathcal{U}(0, |N|) \| q_\sigma(y|x_{un}) \right) \right]$$

Shortcuts-fused Selective Rationalization

SSR

➢ **Data Augmentation**

- **Random Data Augmentation**

We can replace shortcuts tokens with other tokens, sampling randomly from the datastore.

- **Semantic Data Augmentation**

We design a retrieval grounded semantic augmentation method by replacing shortcut tokens with several tokens semantically close to them through retrieval.

Algorithm 1 Semantic Data Augmentation

Input: Supervised dataset $\mathcal{D}_{sup}$, and a well-trained encoder $f_{p_{un}}(\cdot)$ in unsupervised rationalization model $\mathcal{M}_{un}$.
Output: Several semantic related tokens.
Create a global datastore $\mathcal{D}_{global}$:
for j=1 **to** $|\mathcal{D}_{sup}|$ **do**
  Sample x from $\mathcal{D}_{sup}$.
  Construct a key-value pair: (key, value) = $(f_{p_{un}}(x), x)$
end for
$\mathcal{D}_{global} = \{(f_{p_{un}}(x), x), \forall x \in \mathcal{D}_{sup}\}$.
Search the nearest semantic x^r to x in $\mathcal{D}_{global}$, $x^r \neq x$.
Create a local datastore $\mathcal{D}_{local}$:
for i=1 **to** $|x^r|$ **do**
  (key, value) = $(f_{p_{un}}(x_i^r), x_i^r)$.
end for
$\mathcal{D}_{local} = \{(f_{p_{un}}(x_i^r), x_i^r), \forall x_i^r \in x^r\}$.
Search the nearest semantic token x_i^r to x_i in $\mathcal{D}_{local}$, x_i^r does not belong to gold tokens in x or x^r.

Outline

Experiments

Rationale evaluation

- **RQ1: Are the rationales extracted faithful ?**
- **RQ2: How SSR performs as the ground rationales scale changes ?**
- **RQ3: Can our data augmentation methods help existing rationale-based methods improve the performance ?**
- **RQ4: Does SSR capture the faithful rationales for predictions ?**

Experiments

Rationale evaluation

➢ Overall Performance (RQ1)

Table 1: Task F1 and Token F1 of selected rationales for the four dataset. Among them, the underlined scores are the state-of-the-art performances of the supervised rationalization. The results in bold are the best scores in our SSR and its variants.

Methods	Movies		MultiRC		BoolQ		Evidence Inference	
	Task	Token-F1	Task	Token-F1	Task	Token-F1	Task	Token-F1
Vanilla Un-RAT	87.0 ± 0.1	28.1 ± 0.2	57.7 ± 0.4	23.9 ± 0.5	62.0 ± 0.2	19.7 ± 0.4	46.2 ± 0.5	8.9 ± 0.2
IB	84.0 ± 0.0	27.5 ± 0.0	62.1 ± 0.0	24.9 ± 0.0	65.2 ± 0.0	12.8 ± 0.0	46.3 ± 0.0	6.9 ± 0.0
Vanilla Semi-RAT	89.8 ± 0.2	30.4 ± 0.2	63.3 ± 0.4	55.4 ± 0.2	57.3 ± 0.3	43.0 ± 0.1	46.1 ± 0.5	25.1 ± 0.2
IB (25% rationales)	85.4 ± 0.0	28.2 ± 0.0	66.4 ± 0.0	54.0 ± 0.0	63.4 ± 0.0	19.2 ± 0.0	46.7 ± 0.0	10.8 ± 0.0
WSEE	90.1 ± 0.1	32.2 ± 0.1	65.0 ± 0.8	55.8 ± 0.5	59.9 ± 0.4	43.6 ± 0.4	49.2 ± 0.9	14.8 ± 0.8
ST-RAT	87.0 ± 0.0	31.0 ± 0.0	-	-	62.0 ± 0.0	51.0 ± 0.0	46.0 ± 0.0	9.0 ± 0.0
Vanilla Sup-RAT	93.6 ± 0.3	38.2 ± 0.2	63.8 ± 0.2	59.4 ± 0.4	61.5 ± 0.3	51.3 ± 0.2	52.3 ± 0.5	16.5 ± 0.2
Pipeline	86.0 ± 0.0	16.2 ± 0.0	63.3 ± 0.0	41.2 ± 0.0	62.3 ± 0.0	18.4 ± 0.0	70.8 ± 0.0	54.8 ± 0.0
AT-BMC	92.9 ± 0.6	40.2 ± 0.3	65.8 ± 0.2	61.1 ± 0.5	62.1 ± 0.2	52.1 ± 0.2	49.5 ± 0.4	18.6 ± 0.3
SSR_{unif}	94.3 ± 0.3	33.2 ± 0.4	62.8 ± 0.3	56.2 ± 0.2	60.8 ± 0.4	47.6 ± 0.5	46.8 ± 0.3	26.8 ± 0.2
+random DA	90.7 ± 0.3	34.5 ± 0.1	63.6 ± 0.5	56.1 ± 0.3	$\mathbf{61.3 \pm 0.7}$	48.3 ± 0.5	46.0 ± 0.1	33.1 ± 0.2
+semantic DA	90.7 ± 0.2	35.6 ± 0.2	64.7 ± 0.7	42.7 ± 0.4	58.0 ± 0.3	$\mathbf{50.2 \pm 0.3}$	48.7 ± 0.2	33.5 ± 0.4
+mixed DA	$\mathbf{94.5 \pm 0.2}$	35.1 ± 0.1	65.3 ± 0.6	40.3 ± 0.5	60.4 ± 0.2	49.2 ± 0.5	47.6 ± 0.1	$\mathbf{35.2 \pm 0.2}$
−shared W_s and W_p	88.3 ± 0.1	29.8 ± 0.6	60.2 ± 0.3	55.7 ± 0.5	57.4 ± 0.3	43.5 ± 0.2	45.6 ± 0.3	24.9 ± 0.2
SSR_{virt}	90.0 ± 0.0	34.6 ± 0.2	64.2 ± 0.3	$\mathbf{57.0 \pm 0.2}$	58.2 ± 0.5	43.8 ± 0.3	$\mathbf{50.4 \pm 0.3}$	31.3 ± 0.4
+random DA	92.8 ± 0.2	36.7 ± 0.2	65.4 ± 0.2	44.3 ± 0.4	58.3 ± 0.6	47.7 ± 0.3	46.5 ± 0.3	32.4 ± 0.2
+semantic DA	87.6 ± 0.3	36.9 ± 0.1	$\mathbf{66.2 \pm 0.5}$	49.8 ± 0.4	61.1 ± 0.3	48.8 ± 0.2	46.5 ± 0.4	31.1 ± 0.2
+mixed DA	90.5 ± 0.2	$\mathbf{37.4 \pm 0.1}$	64.5 ± 0.6	53.1 ± 0.4	60.3 ± 0.2	49.1 ± 0.1	47.1 ± 0.4	33.4 ± 0.3
−shared W_s and W_p	87.9 ± 0.5	31.9 ± 0.4	62.6 ± 0.3	55.3 ± 0.2	57.6 ± 0.4	43.3 ± 0.5	48.6 ± 0.2	25.8 ± 0.4
−shared W_a and W_p	88.3 ± 0.3	30.4 ± 0.1	62.4 ± 0.9	54.0 ± 2.1	57.5 ± 0.3	42.9 ± 0.1	45.8 ± 0.4	25.0 ± 0.2

IID Dataset

Table 3: Generalization Evaluation on IMDB and SST-2.

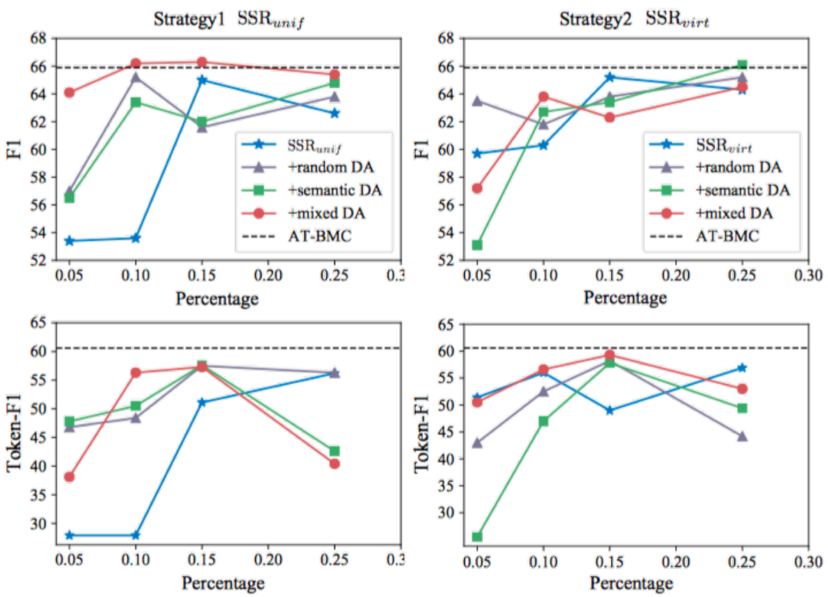
Methods	IMDB	SST-2
Vanilla Un-RAT	85.3 ± 0.2	45.3 ± 8.1
+semantic DA	86.6 ± 0.4	47.8 ± 6.6
Vanilla Semi-RAT	89.5 ± 0.4	75.9 ± 0.7
+semantic DA	89.7 ± 0.3	76.4 ± 0.5
WSEE	90.5 ± 0.3	77.1 ± 0.6
+semantic DA	91.0 ± 0.4	78.3 ± 0.7
SSR_{unif}	90.3 ± 0.2	79.4 ± 0.3
+semantic DA	$\mathbf{90.7 \pm 0.1}$	82.4 ± 0.8
SSR_{virt}	89.9 ± 0.2	79.9 ± 0.4
+semantic DA	90.3 ± 0.3	$\mathbf{83.5 \pm 0.5}$

OOD Dataset

Experiments

Rationale evaluation

➢ Gold Rationale Efficiency (RQ2)



➢ Results of Data Augmentation (RQ3)

Table 2: Task F1 and Token F1 of selected rationales for the four dataset with random DA.

Methods	Movies		MultiRC		BoolQ		Evidence Inference	
	Task	Token-F1	Task	Token-F1	Task	Token-F1	Task	Token-F1
Vanilla Un-RAT	88.0 ± 0.4	28.4 ± 0.3	58.4 ± 0.2	24.7 ± 0.3	62.1 ± 0.3	23.5 ± 0.2	47.0 ± 0.4	10.4 ± 0.3
Vanilla Semi-RAT	90.6 ± 0.3	31.6 ± 0.1	64.2 ± 0.4	56.2 ± 0.3	58.9 ± 0.1	44.5 ± 0.3	45.0 ± 0.3	26.0 ± 0.3
WSEE	89.9 ± 0.4	33.4 ± 0.3	65.3 ± 0.1	55.7 ± 0.3	61.0 ± 0.2	45.5 ± 0.3	50.0 ± 0.3	18.7 ± 0.5
AT-BMC	92.8 ± 0.1	40.4 ± 0.3	66.6 ± 0.6	61.8 ± 0.5	62.0 ± 0.1	52.6 ± 0.2	49.5 ± 0.3	19.4 ± 0.6
SSR$_{unif}$	90.7 ± 0.3	34.5 ± 0.1	63.6 ± 0.5	56.1 ± 0.3	61.3 ± 0.7	48.3 ± 0.5	46.0 ± 0.1	33.1 ± 0.2
SSR$_{virt}$	92.8 ± 0.2	36.7 ± 0.2	65.4 ± 0.2	44.3 ± 0.4	58.3 ± 0.6	47.7 ± 0.3	46.5 ± 0.3	32.4 ± 0.2

Table 3: Task F1 and Token F1 of selected rationales for the four dataset with semantic DA.

Methods	Movies		MultiRC		BoolQ		Evidence Inference	
	Task	Token-F1	Task	Token-F1	Task	Token-F1	Task	Token-F1
Vanilla Un-RAT	88.3 ± 0.2	28.7 ± 0.5	59.0 ± 0.3	25.1 ± 0.4	61.9 ± 0.6	23.7 ± 0.4	47.3 ± 0.3	11.7 ± 0.4
Vanilla Semi-RAT	90.1 ± 0.3	31.3 ± 0.3	64.4 ± 0.1	56.6 ± 0.3	59.1 ± 0.4	44.4 ± 0.2	46.6 ± 0.6	26.9 ± 0.5
WSEE	88.9 ± 0.7	33.1 ± 0.5	64.9 ± 0.3	55.9 ± 0.3	60.9 ± 0.4	46.6 ± 0.6	49.7 ± 0.3	20.9 ± 0.4
AT-BMC	93.2 ± 0.3	40.7 ± 0.5	66.0 ± 0.6	60.9 ± 0.4	62.2 ± 0.3	52.0 ± 0.1	50.0 ± 0.4	22.3 ± 0.6
SSR$_{unif}$	90.7 ± 0.2	35.6 ± 0.2	64.7 ± 0.7	42.7 ± 0.4	58.0 ± 0.3	50.2 ± 0.3	48.7 ± 0.2	33.5 ± 0.4
SSR$_{virt}$	87.6 ± 0.3	36.9 ± 0.1	66.2 ± 0.5	49.8 ± 0.4	61.1 ± 0.3	48.8 ± 0.2	46.5 ± 0.4	31.1 ± 0.2

Experiments

Rationale evaluation

Case Study (RQ4)

Model	Visualized Example	Predicted Label
Vanilla Un-RAT	The film has received a lukewarm response on review sites. What I was in for was a disappointing and overlong film which was anything but the best picture of 1995. What drags it down is its screenplay. It abounds with high production values…	*Negative*
SSR$_{unif}$	The film has received a lukewarm response on review sites. What I was in for was a disappointing and overlong film which was anything but the best picture of 1995. What drags it down is its screenplay. It abounds with high production values…	*Negative*

(a) Visualized selective rationales on Movies. The real label in this case is *Negative*.

Model	Visualized Example	Predicted Label
Vanilla Un-RAT	Mozart is a famous musician and amadeus is a biographical film about him , amadeus is a true work of art . it is one of those few movies of the 80 ' s that will be known for its class , its style , and its intelligence. why is this such a good film...	*Positive*
SSR$_{unif}$	Mozart is a famous musician and amadeus is a biographical film about him , amadeus is a true work of art . it is one of those few movies of the 80 ' s that will be known for its class , its style , and its intelligence. why is this such a good film...	*Positive*

(b) Visualized selective rationales on Movies. The real label in this case is *Positive*.

Model	Visualized Example	Predicted Label
Vanilla Un-RAT	Moonlight mile is replete with acclaimed actors and actresses and tackles a subject that 's potentially moving , the movie is too predictable and too self-conscious to reach a level of high drama.	*Positive*
SSR$_{unif}$	Moonlight mile is replete with acclaimed actors and actresses and tackles a subject that 's potentially moving , the movie is too predictable and too self-conscious to reach a level of high drama.	*Negative*

(c) Visualized selective rationales on SST-2. The real label in this case is *Negative*.

Thank you！