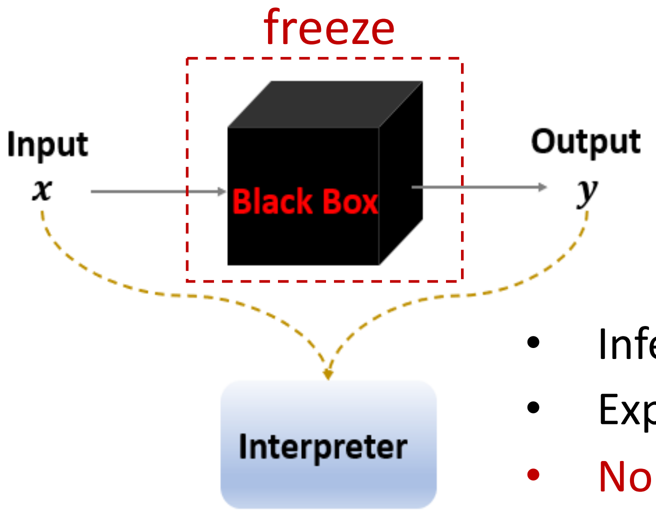
CS 4501/6501 Interpretable Machine Learning

Rationalized Neural Networks

Hanjie Chen, Yangfeng Ji
Department of Computer Science
University of Virginia
{hc9mx, yangfeng}@virginia.edu

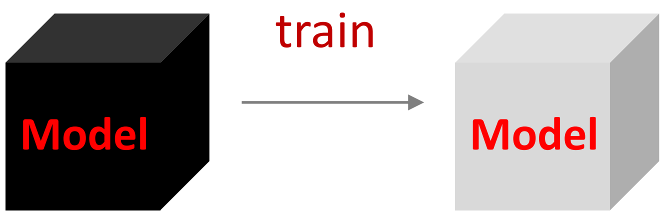
What is the difference?

Explaining a model from the post-hoc manner



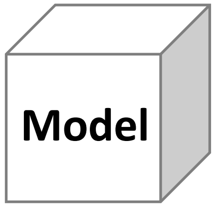
- Inference stage
- Explain model predictions
- No change on model decision making

Improving a model's intrinsic interpretability



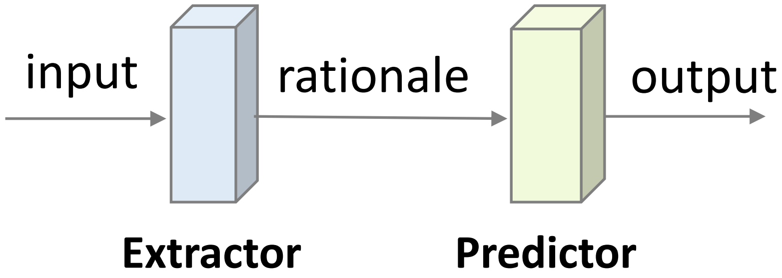
- Training stage
- Make model prediction behavior more interpretable
- No (or minor) change on model architecture

Building Interpretable Neural Network Models



Self-interpretable

Rationalized Neural Networks



Rationalized Neural Networks

- Rationalizing Neural Predictions

- FRESH

Rationalizing Neural Predictions

Tao Lei, Regina Barzilay and Tommi Jaakkola

(EMNLP, 2016)

Rationalizing Neural Predictions

- Rationales: interpretable justifications for model predictions
- Learning problem
  - Prediction
  - Rationale generation

Rationalizing Neural Predictions

- Rationales: interpretable justifications for model predictions
- Learning problem
  - Prediction
  - Rationale generation

    (subsets of words extracted from the input)

- short and coherent pieces of text (e.g., phrases)
- suffice for prediction

Review

the beer was n't what i expected, and i'm not sure it's "true to style", but i thought it was delicious. **a very pleasant ruby red-amber color** with a relatively brilliant finish, but a limited amount of carbonation, from the look of it. aroma is what i think an amber ale should be - a nice blend of caramel and happiness bound together.

Ratings *Look:* 5 stars *Smell:* 4 stars

Extractive Rationale Generation

A sequence of words	generator	Rationale	encoder	Prediction
$\boldsymbol{x} = \{x_1, x_2, \cdots, x_l\}$	$\longrightarrow$	$gen(\boldsymbol{x})$	$\longrightarrow$	$enc(gen(\boldsymbol{x}))$
		short and sufficient		$\approx enc(\boldsymbol{x})$

Extractive Rationale Generation

A sequence of words $\quad$ generator $\quad$ Rationale $\quad$ encoder $\quad$ Prediction

$$\boldsymbol{x} = \{x_1, x_2, \cdots, x_l\} \longrightarrow gen(\boldsymbol{x}) \longrightarrow enc(gen(\boldsymbol{x}))$$

short and sufficient

$\approx enc(\boldsymbol{x})$

$gen(\cdot)$: a tagging model

$$\{0,1,\cdots,0\}_l$$

Extractive Rationale Generation

A sequence of words

$\boldsymbol{x} = \{x_1, x_2, \cdots, x_l\}$ $\xrightarrow{\text{generator}}$ Rationale $gen(\boldsymbol{x})$ $\xrightarrow{\text{encoder}}$ Prediction $enc(gen(\boldsymbol{x}))$

Train jointly

short and sufficient

No additional supervision

$\approx enc(\boldsymbol{x})$

$gen(\cdot)$: a tagging model

$\{0,1,\cdots,0\}_l$

Encoder and Generator

Encoder $\boldsymbol{enc}(\cdot)$

$$\tilde{y} = enc(\boldsymbol{x})$$

$$\mathcal{L}(\boldsymbol{x}, y) = \|\tilde{y} - y\|_2^2 = \|enc(\boldsymbol{x}) - y\|_2^2$$

Encoder and Generator

Generator $\boldsymbol{gen}(\cdot)$

$$gen(\boldsymbol{x}) \longrightarrow \boldsymbol{z} = \{z_1, z_2, \cdots, z_l\} \quad z_t \in \{0, 1\}$$

$$\boldsymbol{z} \sim gen(\boldsymbol{x}) \equiv p(\boldsymbol{z}|\boldsymbol{x})$$

Encoder and Generator

Generator $\boldsymbol{gen}(\cdot)$

$$gen(\boldsymbol{x}) \longrightarrow \boldsymbol{z} = \{z_1, z_2, \cdots, z_l\} \quad z_t \in \{0, 1\}$$

$$\boldsymbol{z} \sim gen(\boldsymbol{x}) \equiv p(\boldsymbol{z}|\boldsymbol{x})$$

$$p(\boldsymbol{z}|\boldsymbol{x}) = \prod_{t=1}^{l} p(z_t|\boldsymbol{x}) \quad \text{(independent selection)}$$

Encoder and Generator

Generator $\boldsymbol{gen}(\cdot)$

$$gen(\boldsymbol{x}) \longrightarrow \boldsymbol{z} = \{z_1, z_2, \cdots, z_l\} \quad z_t \in \{0,1\}$$

$$\boldsymbol{z} \sim gen(\boldsymbol{x}) \equiv p(\boldsymbol{z}|\boldsymbol{x})$$

$$p(\boldsymbol{z}|\boldsymbol{x}) = \prod_{t=1}^{l} p(z_t|\boldsymbol{x}) \quad \text{(independent selection)}$$

Or

$$p(\boldsymbol{z}|\boldsymbol{x}) = \prod_{t=1}^{l} p(z_t|\boldsymbol{x}, z_1 \cdots, z_{t-\mathbf{1}}) \quad \text{(recurrent selection)}$$

Encoder and Generator

Generator $\boldsymbol{gen}(\cdot)$

$$gen(\boldsymbol{x}) \longrightarrow \boldsymbol{z} = \{z_1, z_2, \cdots, z_l\} \quad z_t \in \{0, 1\}$$

$$\boldsymbol{z} \sim gen(\boldsymbol{x}) \equiv p(\boldsymbol{z}|\boldsymbol{x})$$

$$p(\boldsymbol{z}|\boldsymbol{x}) = \prod_{t=1}^{l} p(z_t|\boldsymbol{x}) \quad \text{(independent selection)}$$

Or

$$p(\boldsymbol{z}|\boldsymbol{x}) = \prod_{t=1}^{l} p(z_t|\boldsymbol{x}, z_1 \cdots, z_{t-\mathbf{1}}) \quad \text{(recurrent selection)}$$

The component distributions are modeled via a shared bi-directional recurrent neural network

Encoder and Generator

Joint objective

A rationale $(\boldsymbol{z}, \boldsymbol{x})$ corresponds to the selected words, i.e., $\{x_t | z_t = 1\}$

The rationale should suffice as a replacement for the input text:

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{x}, y) = \| enc(\boldsymbol{z}, \boldsymbol{x}) - y \|_2^2$$

Encoder and Generator

Joint objective

A rationale $(\boldsymbol{z}, \boldsymbol{x})$ corresponds to the selected words, i.e., $\{x_t | z_t = 1\}$

The rationale should suffice as a replacement for the input text:

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{x}, y) = \| enc(\boldsymbol{z}, \boldsymbol{x}) - y \|_2^2$$

The loss function depends directly on the encoder but only indirectly on the generator via the sampled selection

Encoder and Generator

Joint objective

A rationale $(\boldsymbol{z}, \boldsymbol{x})$ corresponds to the selected words, i.e., $\{x_t | z_t = 1\}$

The rationale should suffice as a replacement for the input text:

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{x}, y) = \| enc(\boldsymbol{z}, \boldsymbol{x}) - y \|_2^2$$

The rationale should be short and coherent:

(A few and consecutive words, e.g., phrases)

$$\Omega(\boldsymbol{z}) = \lambda_1 \|\boldsymbol{z}\| + \lambda_2 \sum_t |z_t - z_{t-1}|$$

(Control the number of selections)

(Encourage the continuity of selections)

Encoder and Generator

Joint objective

A rationale $(\boldsymbol{z}, \boldsymbol{x})$ corresponds to the selected words, i.e., $\{x_t | z_t = 1\}$

The rationale should suffice as a replacement for the input text:

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{x}, y) = \|enc(\boldsymbol{z}, \boldsymbol{x}) - y\|_2^2$$

The rationale should be short and coherent:

(A few and consecutive words, e.g., phrases)

Objective

$\mathcal{L}(\boldsymbol{z}, \boldsymbol{x}, y) + \Omega(\boldsymbol{z})$

$$\Omega(\boldsymbol{z}) = \lambda_1 \|\boldsymbol{z}\| + \lambda_2 \sum_t |z_t - z_{t-1}|$$

Question?

Experiments

Multi-aspect Sentiment Analysis

Dataset: BeerAdvocate review (McAuley et al., 2012)

- 1.5 million reviews written by the website users
- the reviews are naturally multi-aspect
- each of them contains multiple sentences
- describing the overall impression
- one particular aspect of a beer (appearance, smell, palate, taste)
- an overall score ([0, 1]) and the score for each aspect
- Sentence-level annotations: indicating what aspect a sentence covers

Experiments

Multi-aspect Sentiment Analysis

Assessing different neural encoder architectures

	D	d	l	$\|\theta\|$	MSE
SVM	260k	-	-	2.5M	0.0154
SVM	1580k	-	-	7.3M	0.0100
LSTM	260k	200	2	644k	0.0094
RCNN	260k	200	2	323k	**0.0087**

(recurrent convolutional
neural networks)

Experiments

Multi-aspect Sentiment Analysis

Assessing different neural encoder architectures

	D	d	l	$\|\theta\|$	MSE
SVM	260k	-	-	2.5M	0.0154
SVM	1580k	-	-	7.3M	0.0100
LSTM	260k	200	2	644k	0.0094
RCNN	260k	200	2	323k	**0.0087**

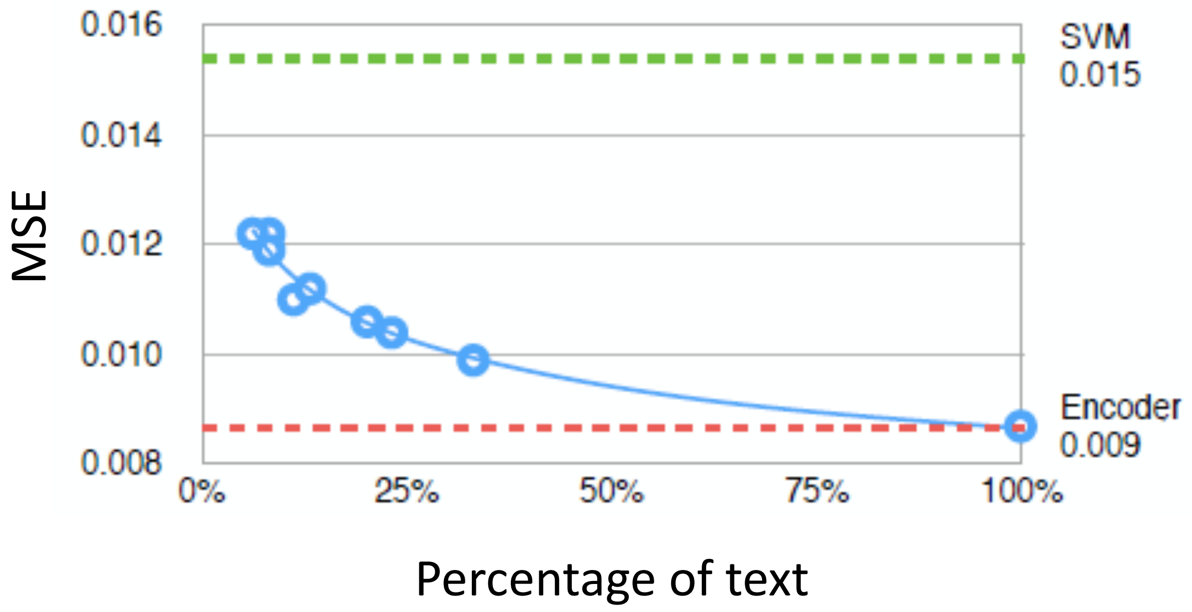
(recurrent convolutional neural networks)

The generator is also constructed with RCNN units

Experiments

Multi-aspect Sentiment Analysis

Prediction performance

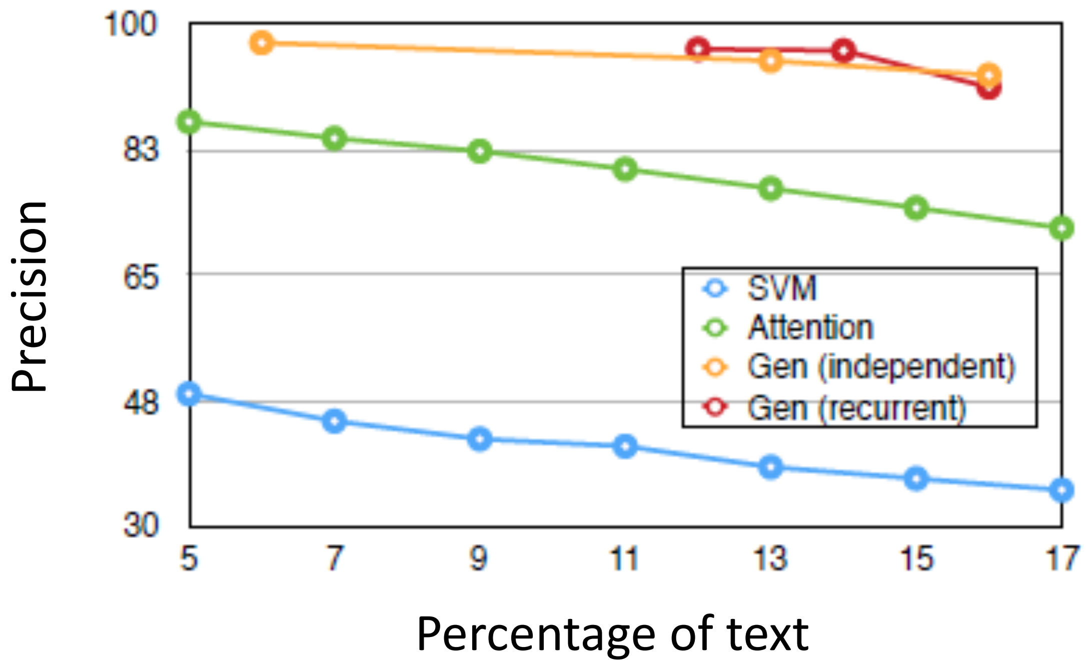


Sacrifice of performance

Experiments

Multi-aspect Sentiment Analysis

Rationale selection



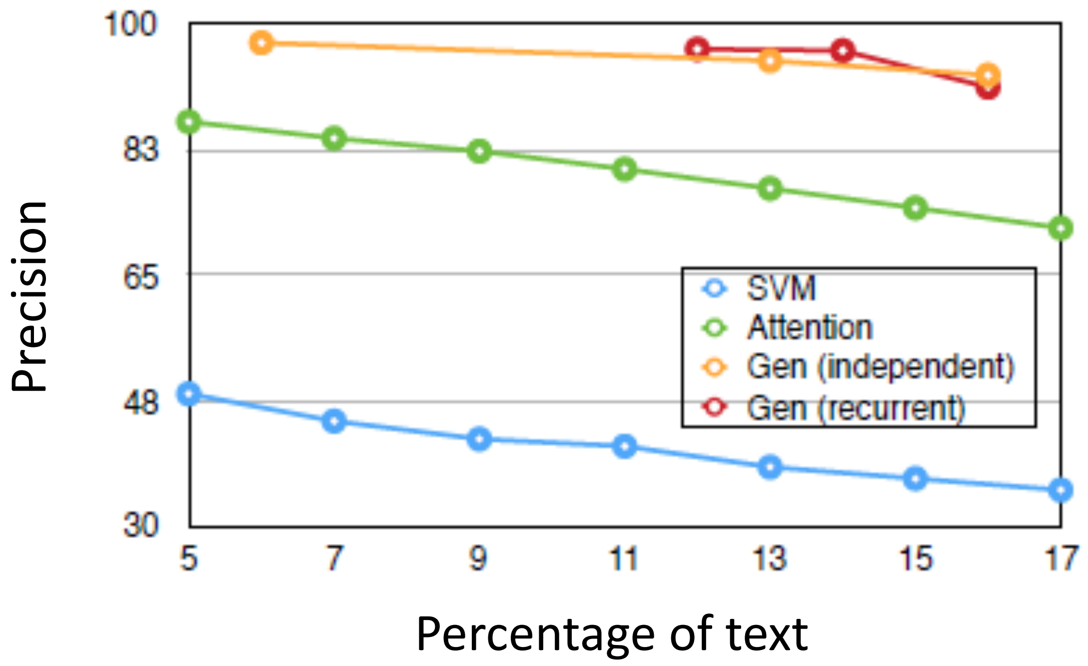
SVM successively extracts unigram or bigram with the highest feature

The attention-based model selects words based on their attention weights

Experiments

Multi-aspect Sentiment Analysis

Rationale selection



✓ The encoder-generator network extracts text pieces describing the target aspect with high precision

Experiments

Multi-aspect Sentiment Analysis

Rationale selection (appearance, smell, palate)

a beer that is not sold in my neck of the woods , but managed to get while on a roadtrip . poured into an imperial pint glass with a generous head that sustained life throughout . nothing out of the ordinary here , but a good brew still . body was kind of heavy , but not thick . the hop smell was excellent and enticing . very drinkable

---

very dark beer . pours a nice finger and a half of creamy foam and stays throughout the beer . smells of coffee and roasted malt . has a major coffee-like taste with hints of chocolate . if you like black coffee , you will love this porter . creamy smooth mouthfeel and definitely gets smoother on the palate once it warms . it 's an ok porter but i feel there are much better one 's out there .

---

i really did not like this . it just seemed extremely watery . i dont ' think this had any carbonation whatsoever . maybe it was flat , who knows ? but even if i got a bad brew i do n't see how this would possibly be something i 'd get time and time again . i could taste the hops towards the middle , but the beer got pretty nasty towards the bottom . i would never drink this again , unless it was free . i 'm kind of upset i bought this .

---

a : poured a nice dark brown with a tan colored head about half an inch thick , nice red/garnet accents when held to the light . little clumps of lacing all around the glass , not too shabby . not terribly impressive though s : smells like a more guinness-y guinness really , there are some roasted malts there , signature guinness smells , less burnt though , a little bit of chocolate ... ... m : relatively thick , it is n't an export stout or imperial stout , but still is pretty hefty in the mouth , very smooth , not much carbonation . not too shabby d : not quite as drinkable as the draught , but still not too bad . i could easily see drinking a few of these .

Question?

Rationalized Neural Networks

- Rationalizing Neural Predictions
- FRESH

Learning to Faithfully Rationalize by Construction

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, Byron C. Wallace

(ACL, 2020)

Key Property

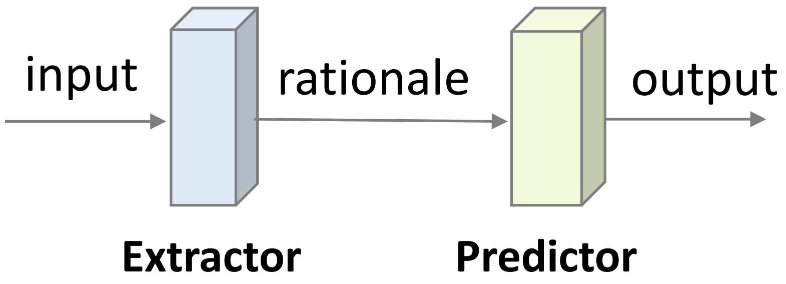
Faithfulness: an explanation provided by a model is faithful if it reflects the information actually used by said model to come to a disposition

(Lipton, 2018)

Problem

(Lei et al., 2016)

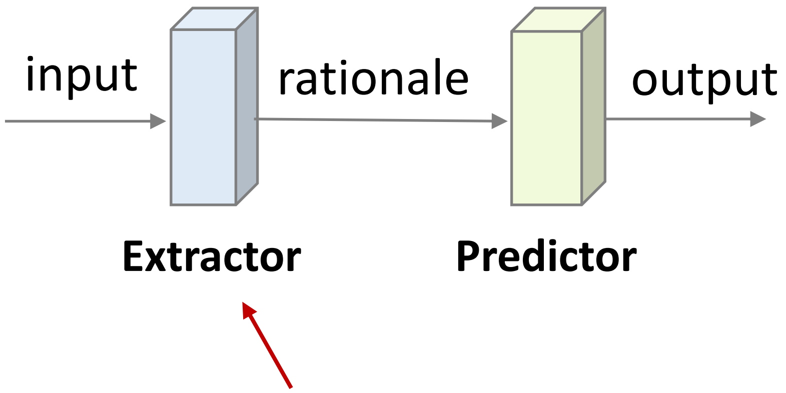
The difficulty of training the two components jointly under only instance-level supervision



Problem

(Lei et al., 2016)

The difficulty of training the two components jointly under only instance-level supervision

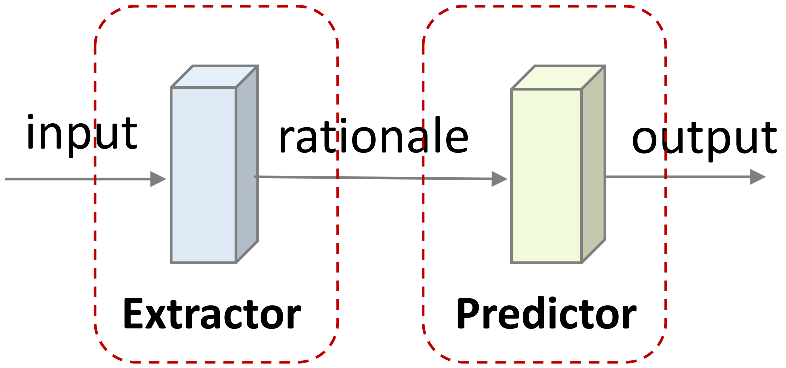


No supervision
(e.g., token labels)

The discrete selection over input tokens complicates training, leading to high variance and requiring careful hyperparameter tuning

FRESH

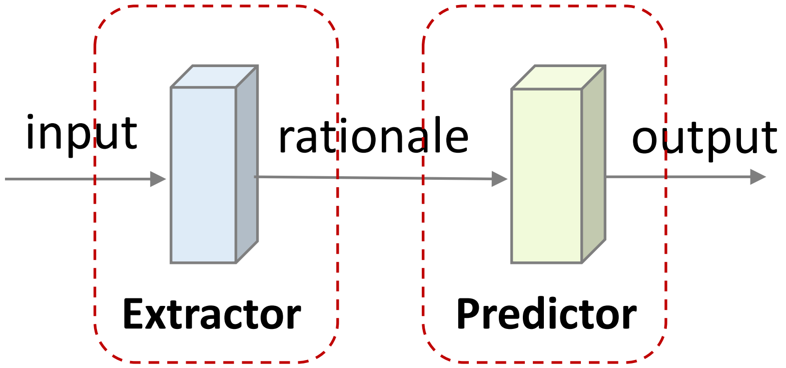
Faithful Rationale Extraction from Saliency tHresholding (FRESH)



Train separately

FRESH

Faithful Rationale Extraction from Saliency tHresholding (FRESH)

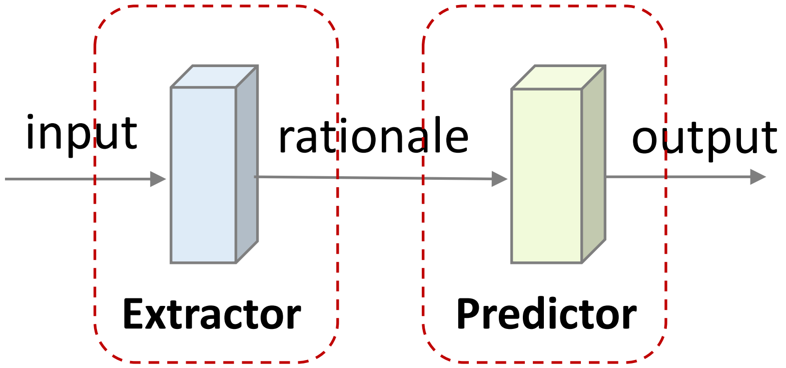


Train separately

FRESH is faithful by construction: the snippet that is ultimately used to inform a prediction can be presented as a faithful explanation

FRESH

Faithful Rationale Extraction from Saliency tHresholding (FRESH)



Train separately

FRESH is plausible: the extracted rationales are intuitive to humans

FRESH

End-to-End Rationale Extraction

Text classification task

n input documents $\{x_1, x_2, \cdots, x_n\}$

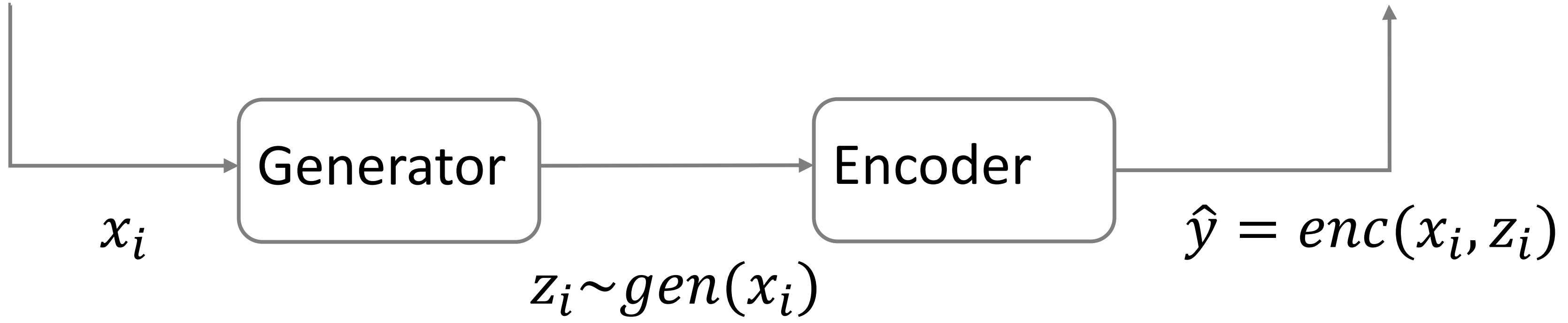
Assigned labels $\{y_1, y_2, \cdots, y_n\}$

FRESH

End-to-End Rationale Extraction

Text classification task

n input documents $\{x_1, x_2, \cdots, x_n\}$ Assigned labels $\{y_1, y_2, \cdots, y_n\}$



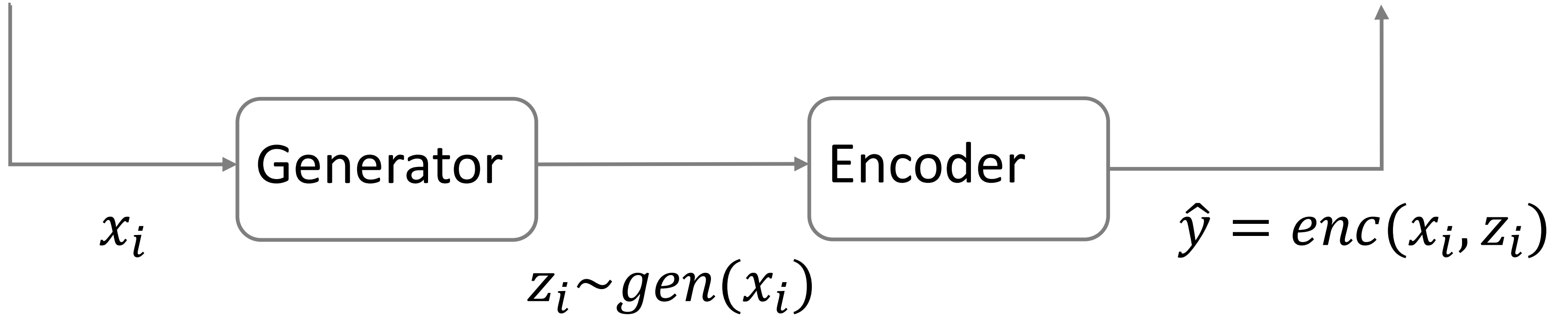
FRESH

End-to-End Rationale Extraction

Text classification task

n input documents $\{x_1, x_2, \cdots, x_n\}$

Assigned labels $\{y_1, y_2, \cdots, y_n\}$



Objective

$$\min_{\theta_{enc}, \theta_{gen}} \sum_{i=1}^{n} E_{z_i \sim gen(x_i)} \mathcal{L}(enc(x_i, z_i), y_i)$$

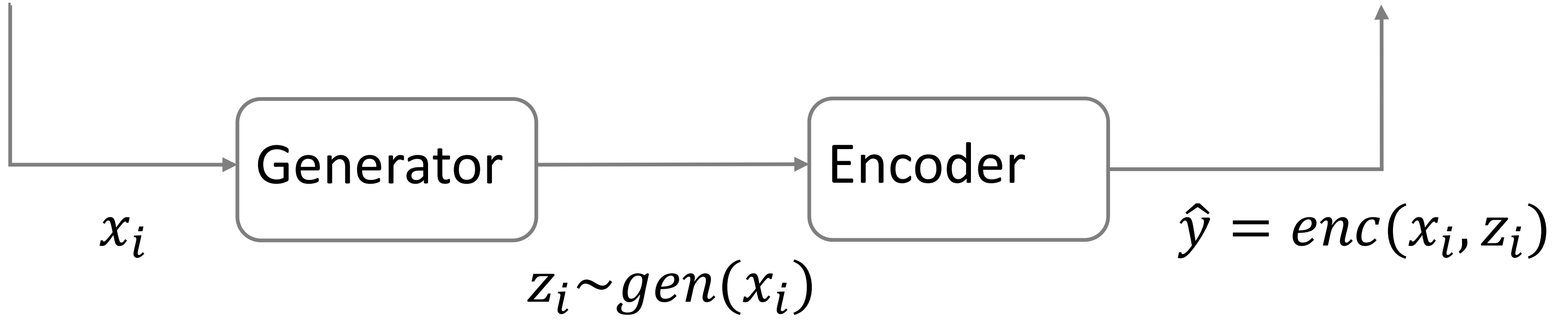
FRESH

End-to-End Rationale Extraction

Text classification task

n input documents $\{x_1, x_2, \cdots, x_n\}$

Assigned labels $\{y_1, y_2, \cdots, y_n\}$



Objective

$$\min_{\theta_{enc}, \theta_{gen}} \sum_{i=1}^{n} E_{z_i \sim gen(x_i)} \mathcal{L}(enc(x_i, z_i), y_i)$$

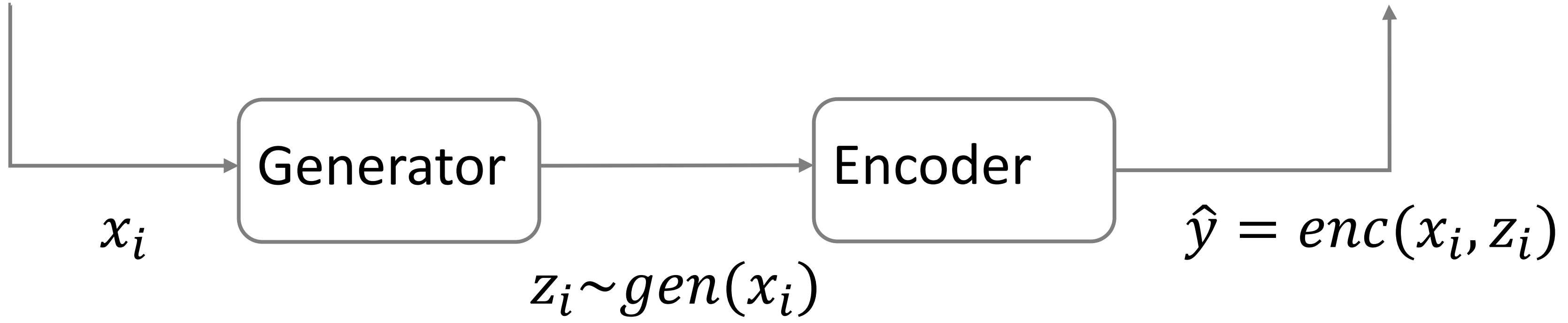
Marginalizing over all possible rationales z causes difficulty in optimization

FRESH

End-to-End Rationale Extraction

Text classification task

n input documents $\{x_1, x_2, \cdots, x_n\}$ Assigned labels $\{y_1, y_2, \cdots, y_n\}$



Objective

$$\min_{\theta_{enc}, \theta_{gen}} \sum_{i=1}^{n} E_{z_i \sim gen(x_i)} \mathcal{L}(enc(x_i, z_i), y_i)$$

Conciseness and contiguity $\Omega(\boldsymbol{z}) = \lambda_1 max\left(0, \frac{|z|}{L} - d\right) + \lambda_2 \sum_t \frac{|z_t - z_{t-1}|}{L-1}$

Question?

FRESH

Three independent components

Support model (supp)

Extractor model (ext)

Classifier (pred)

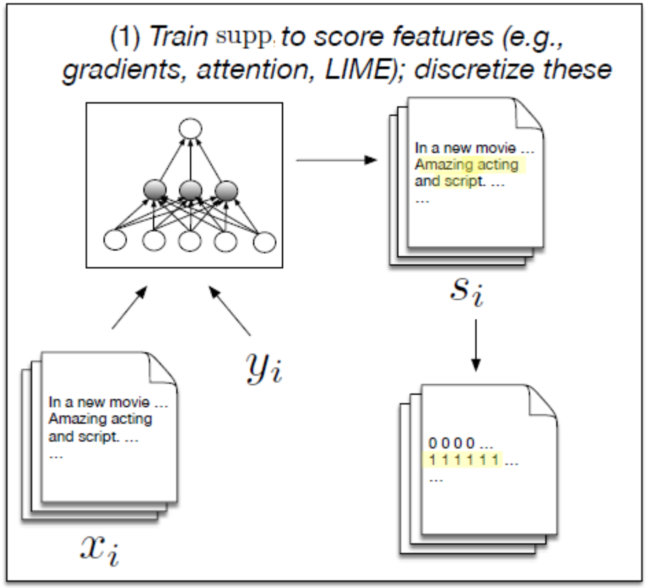
FRESH

Three independent components

Support model
(supp)

Extractor model
(ext)

Classifier
(pred)



- Train supp end-to-end to predict y
- Use its outputs only to extract continuous feature importance scores

(post-hoc explanations)

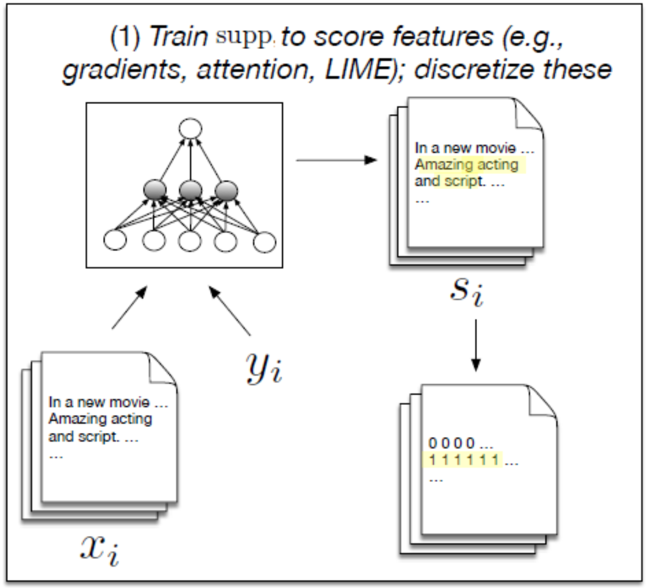
FRESH

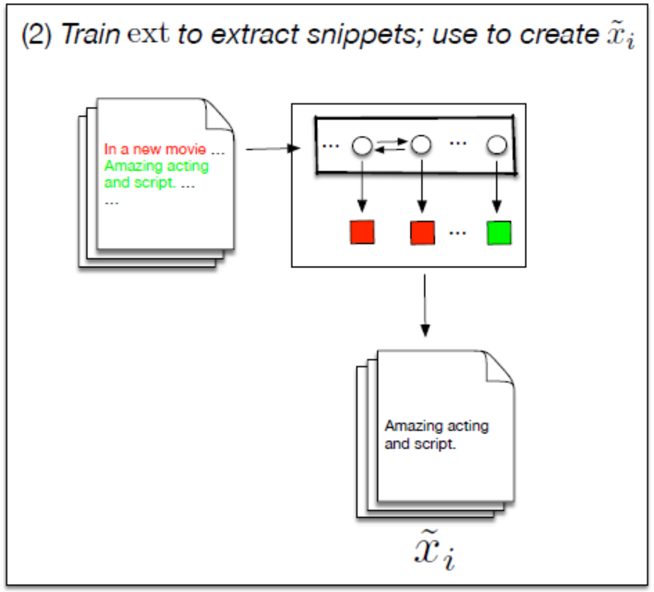
Three independent components

Support model (supp)

Extractor model (ext)

Classifier (pred)





- Use the importance scores to train ext (e.g., treating the top k tokens as the tartget rationale)
- Extract snippets

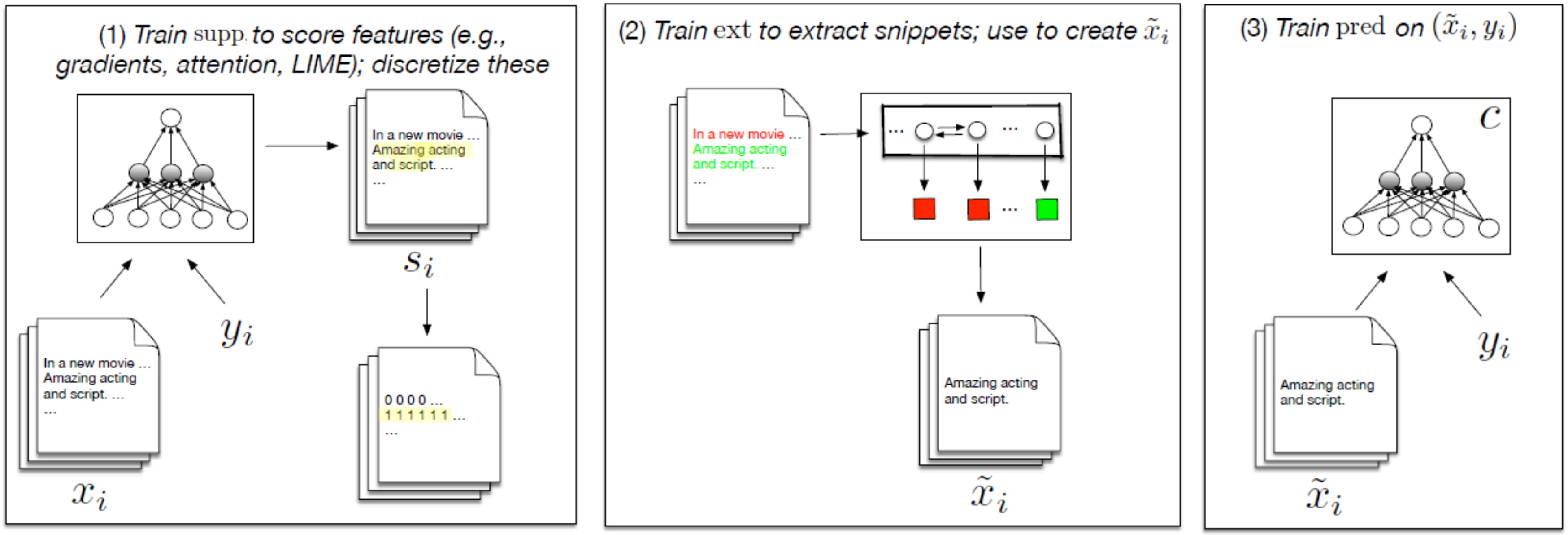
FRESH

Three independent components

Support model (supp)

Extractor model (ext)

Classifier (pred)



- Train pred on the extracted snippets

FRESH

Implementation

Support model (supp)

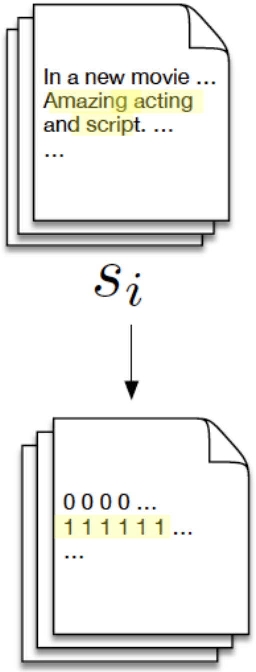
Extractor model (ext)

Classifier (pred)

Support model (supp) → BERT encoder → Attention/gradient-based importance scores

Discretizing Soft Scores

- Contiguous: select the span of length k with the highest total score
- Top-k (non-contiguous): select top-k individual tokens



FRESH

Implementation

Support model (supp) → BERT encoder → Attention/gradient-based importance scores

Extractor model (ext) → discretization heuristics or BERT for sequencing tagging

Classifier (pred)

FRESH

Implementation

Support model (supp) → BERT encoder → Attention/gradient-based importance scores

Extractor model (ext) → discretization heuristics or BERT for sequencing tagging

Classifier (pred) → BERT for classification

FRESH

Implementation

Support model (supp)

BERT encoder

Attention/gradient-based importance scores

Extractor model (ext)

discretization heuristics

or

BERT for sequencing tagging

Classifier (pred)

BERT for classification

Leverage post-hoc explanations to guide rationale extraction

Question?

Experiments

Empirical results (Lei et al., 2016)

- Hyperparameter sensitivity

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{x}, y) = \|enc(\boldsymbol{z}, \boldsymbol{x}) - y\|_2^2$$

$$\Omega(\boldsymbol{z}) = \lambda_1 \|\boldsymbol{z}\| + \lambda_2 \sum_t |z_t - z_{t-1}|$$

- Model performance is sensitive to hyperparameters (λ_1, λ_2)
- Hyperparameter search is time-consuming

Experiments

Empirical results (Lei et al., 2016)

- Hyperparameter sensitivity

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{x}, y) = \| enc(\boldsymbol{z}, \boldsymbol{x}) - y \|_2^2$$

$$\Omega(\boldsymbol{z}) = \lambda_1 \|\boldsymbol{z}\| + \lambda_2 \sum_t |z_t - z_{t-1}|$$

- Model performance is sensitive to hyperparameters (λ_1, λ_2)
- Hyperparameter search is time-consuming

- High variance in performance

Performance varies across different random seeds

Experiments

Prediction performance

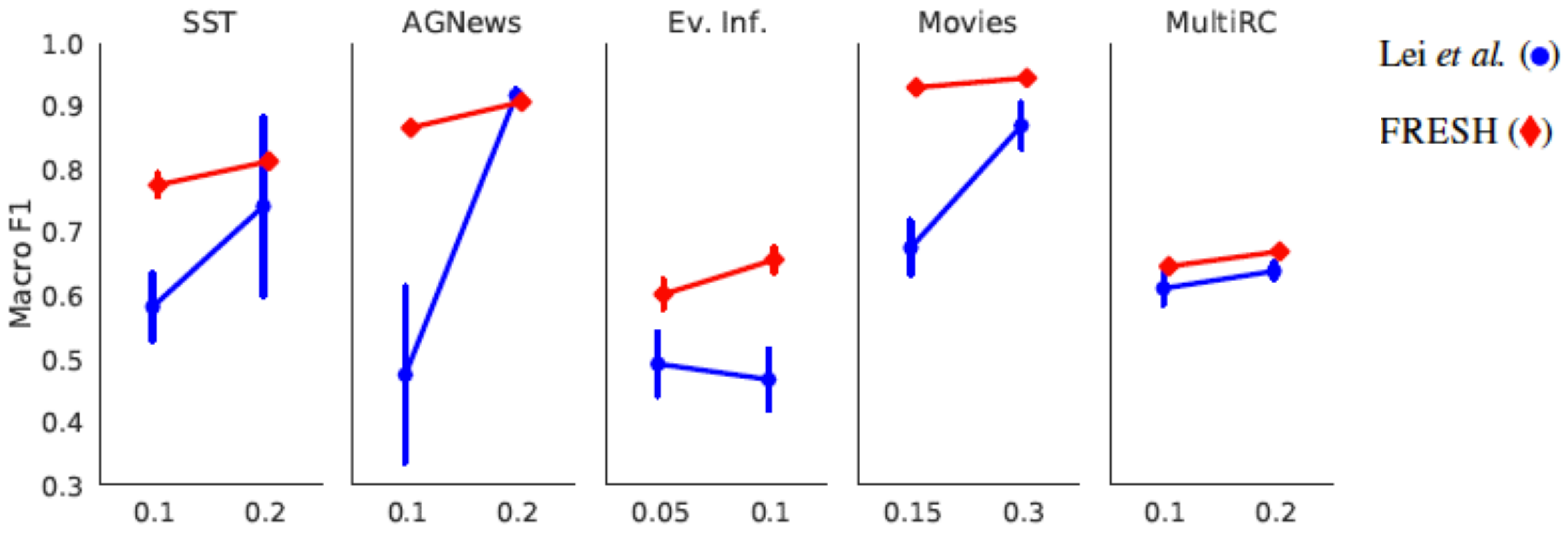
- Outperform baseline methods
- Performance drops compared with the baseline with full text as input

Saliency	Rationale	SST (20%)	AGNews (20%)	Ev. Inf. (10%)	Movies (30%)	MultiRC (20%)
Full text	–	.90 (.89-.90)	.94 (.94-.94)	.73 (.73-.78)	.95 (.93-.97)	.68 (.68-.69)
Lei *et al.*	contiguous	.71 (.49-.83)	.87 (.85-.89)	.53 (.45-.56)	.83 (.80-.92)	.62 (.62-.64)
	top k	.74 (.47-.84)	**.92 (.90-.92)**	.47 (.38-.53)	.87 (.80-.91)	.64 (.61-.65)
Bastings *et al.*	contiguous	.60 (.58-.62)	.77 (.18-.78)	.45 (.40-.49)	—	.41 (.30-.50)
	top k	.59 (.58-.61)	.72 (.19-.80)	.50 (.38-.60)	—	.44 (.30-.55)
Gradient	contiguous	.70 (.69-.72)	.85 (.84-.85)	.67 (.62-.68)	**.94 (.92-.95)**	**.67 (.66-.67)**
	top k	.68 (.67-.70)	.86 (.85-.86)	.62 (.61-.64)	.93 (.92-.94)	.66 (.65-.67)
[CLS] Attn	contiguous	**.81 (.80-.82)**	.88 (.88-.89)	**.68 (.59-.73)**	.93 (.90-.94)	.63 (.60-.62)
	top k	**.81 (.80-.82)**	**.91 (.90-.91)**	.66 (.64-.70)	**.94 (.93-.95)**	.63 (.62-.64)

Experiments

Varying rationale length

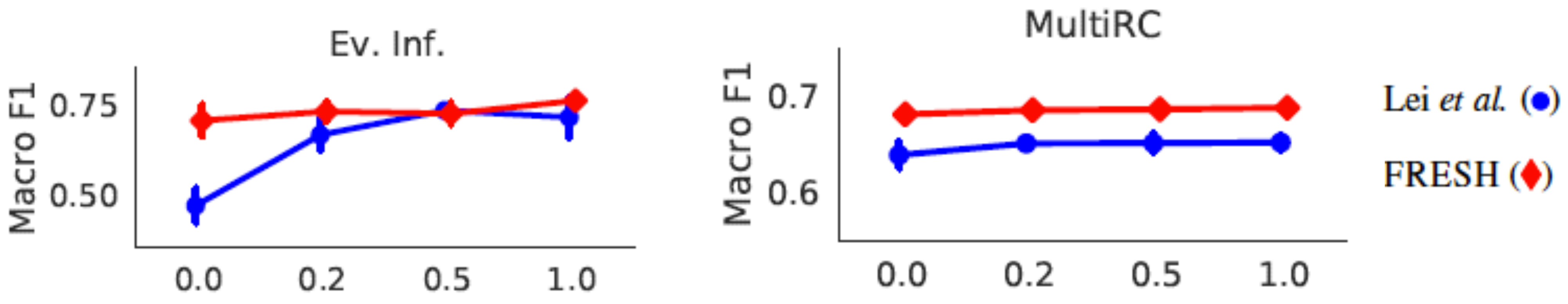
The effectiveness of FRESH even in constrained settings



Experiments

Incorporating human rationale supervision

- Varying amounts of rationale-level supervision (0, 20%, 50%, 100%)
- Introducing an additional binary cross entropy term into the objective
- Overall, mixing in rationale-level supervision can improve performance (not much)



Human Analysis

Sufficiency: Can a human predict the correct label given only the rationale?

Readability and understandability: test the user's preference for a certain style of rationale beyond their ability to predict the correct label
(one hypothesis is that humans will prefer contiguous to non-contiguous rationales)

Human Analysis

Sufficiency: Can a human predict the correct label given only the rationale?

Readability and understandability: test the user's preference for a certain style of rationale beyond their ability to predict the correct label
(one hypothesis is that humans will prefer contiguous to non-contiguous rationales)

FRESH rationales (both contiguous and noncontiguous)

Baselines:

- Human rationales
- Randomly selected "rationales" of length k
- Rationales from Lei et al., 2016 models

Human Analysis

Rationales →  →

- Classify examples
- Rate their confidence (1-4)
- Rate how easy the text is to read and understand (1-5)

Human Analysis

Rationales →

→

- Classify examples
- Rate their confidence (1-4)
- Rate how easy the text is to read and understand (1-5)

Rationale Source	Human Acc.	Confidence (1–4)	Readability (1–5)
Human	.99	3.44 ±0.53	3.82 ±0.56
Random			
Contiguous	.84	3.18 ±0.55	3.80 ±0.57
Non-Contiguous	.65	2.09 ±0.51	2.07 ±0.69
Lei et al. 2016			
Contiguous	.88	3.39 ±0.48	4.17 ±0.59
Non-Contiguous	.84	2.97 ±0.72	2.90 ±0.88
FRESH Best			
Contiguous	.92	3.31 ±0.48	3.88 ±0.57
Non-Contiguous	.87	3.23 ±0.47	3.63 ±0.59

- Humans achieve the best performance on FRESH rationales
- Humans exhibit a strong preference for contiguous rationales

Question?

Reference

- Lei, Tao, Regina Barzilay, and Tommi Jaakkola. "Rationalizing neural predictions." *arXiv preprint arXiv:1606.04155* (2016).
- Jain, Sarthak, et al. "Learning to faithfully rationalize by construction." *arXiv preprint arXiv:2005.00115* (2020).