CS 4501/6501 Interpretable Machine Learning

Improving neural network intrinsic interpretability

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

What is interpretability?

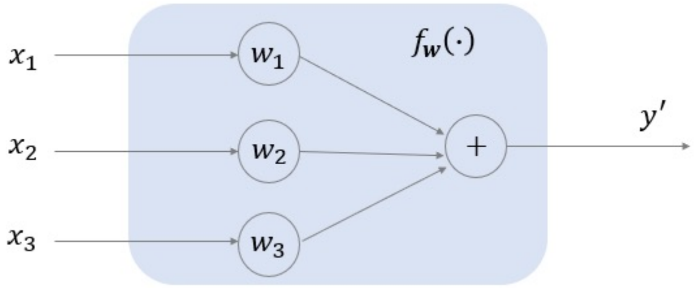
There is no standard or mathematical definition of interpretability

- Interpretability is the degree to which a human can understand the cause of a decision [Miller, 2019]

- Interpretability is the degree to which a human can consistently predict the model's result [Kim et al., 2016]

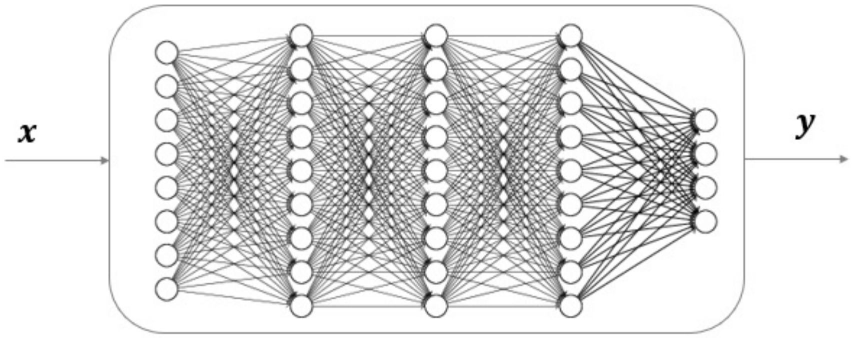
What is intrinsic interpretability?

A simple model is usually more interpretable than a complex neural network model



- Three parameters (w_1, w_2, w_3)
- $y' = w_1 x_1 + w_2 x_2 + w_3 x_3$
- Contributions:

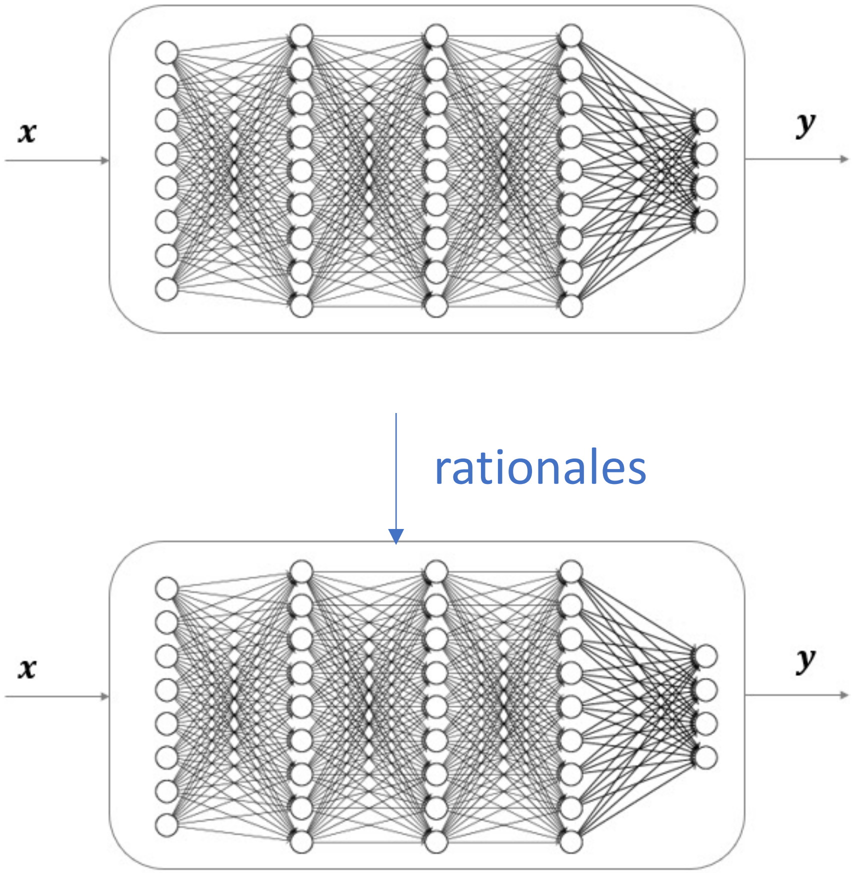
$$x_1 : w_1 x_1$$
$$x_2 : w_2 x_2$$
$$x_3 : w_3 x_3$$



- Millions of parameters
- $\boldsymbol{y}' = f_{\boldsymbol{w}}(\boldsymbol{x})$ (complex transformations)
- Model decision-making and feature attributions are unclear

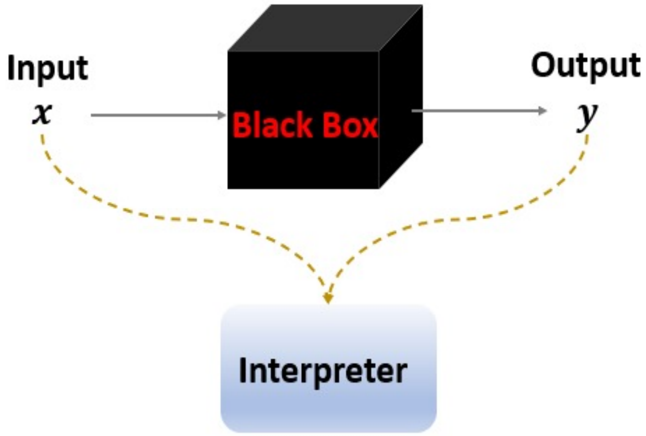
What is intrinsic interpretability?

Similar models trained in different ways may have different interpretability

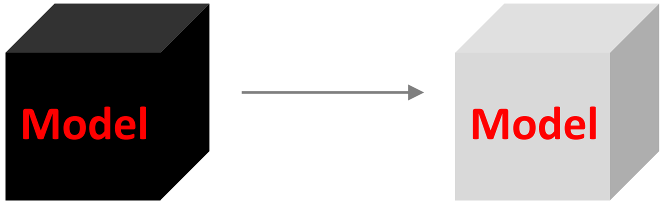


What is the difference?

Explaining a model from the post-hoc manner

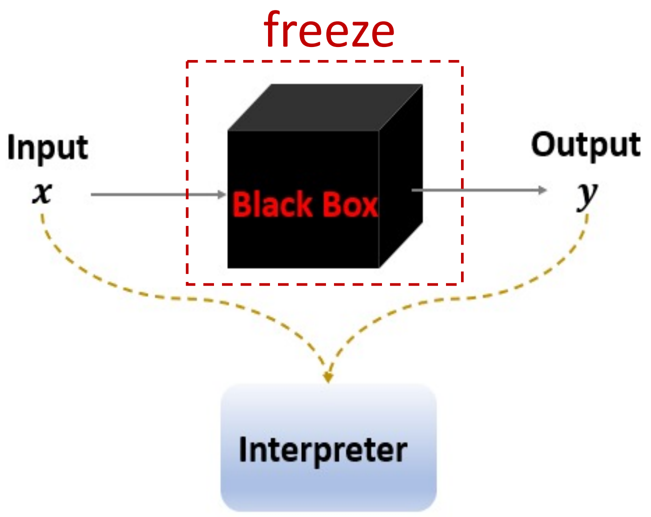


Improving a model's intrinsic interpretability



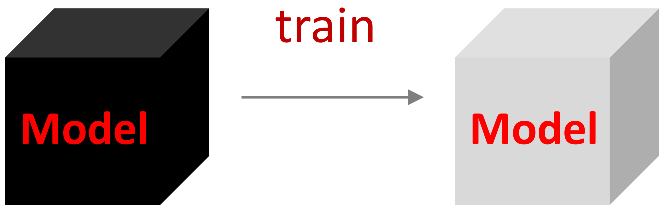
What is the difference?

Explaining a model from the post-hoc manner



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

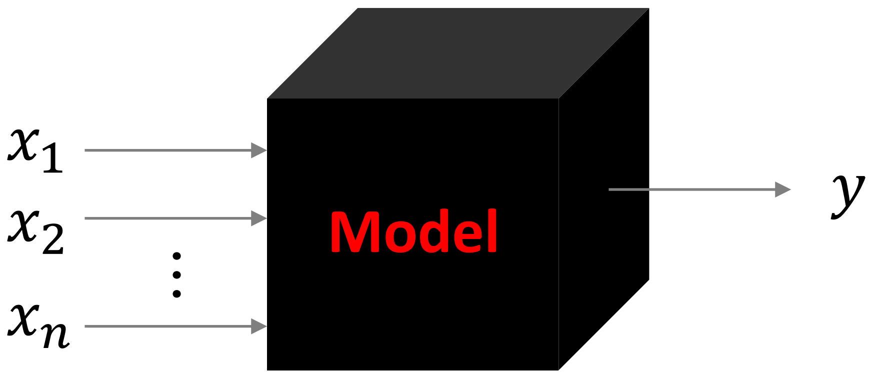
Improving a model's intrinsic interpretability

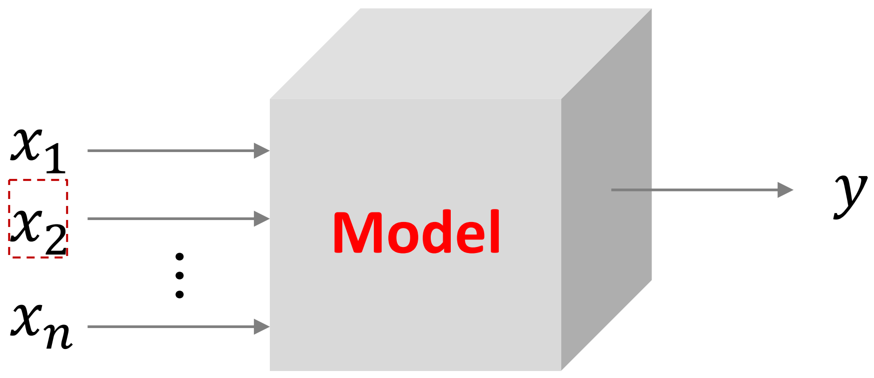


- Training stage
- Make model prediction behavior more interpretable
- No (or minor) change on model architecture

How to improve model intrinsic interpretability?

Teach the model to focus on important features to make predictions





Improving Intrinsic Interpretability

- Training with rationales

- Variational word masks (VMASK)

e-SNLI: Natural Language Inference with Natural Language Explanations

Oana-Maria Camburu, Tim Rocktäschel,
Thomas Lukasiewicz, Phil Blunsom

(NeurIPS, 2018)

e-SNLI

- An extension of the Stanford Natural Language Inference (SNLI) dataset

[Bowman et al., 2015]

- With human-annotated natural language explanations of the entailment relations

- Incorporating these explanations into model training for improving model interpretability

e-SNLI

- **Task**: Natural Language Inference (NLI)

  Predict the relationship between a premise and a hypothesis as "entailment", "contradiction", or "neutral"

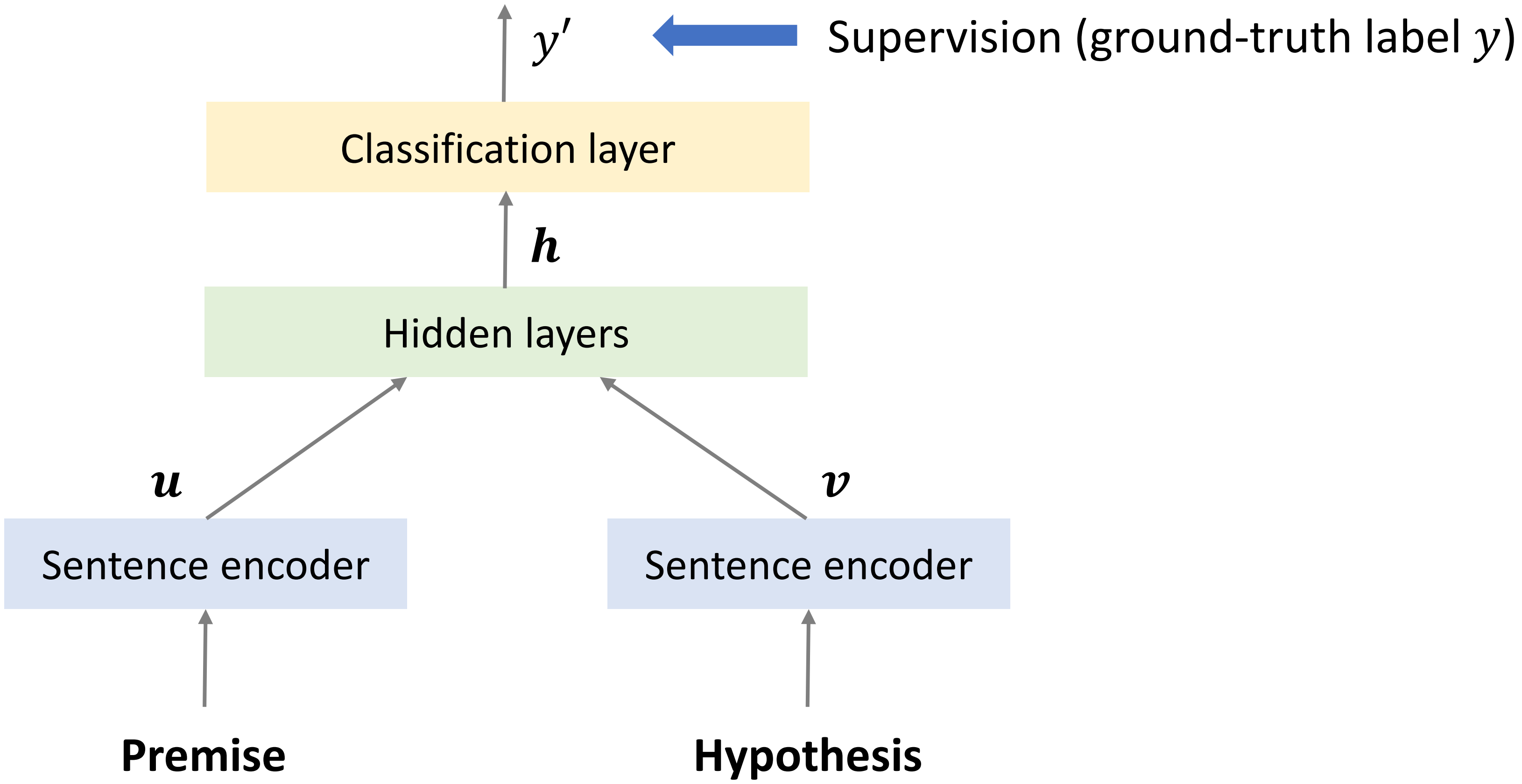
Label: entailment

Premise: A runner wearing purple strives for the finish line

Hypothesis: A runner wants to head for the finish line

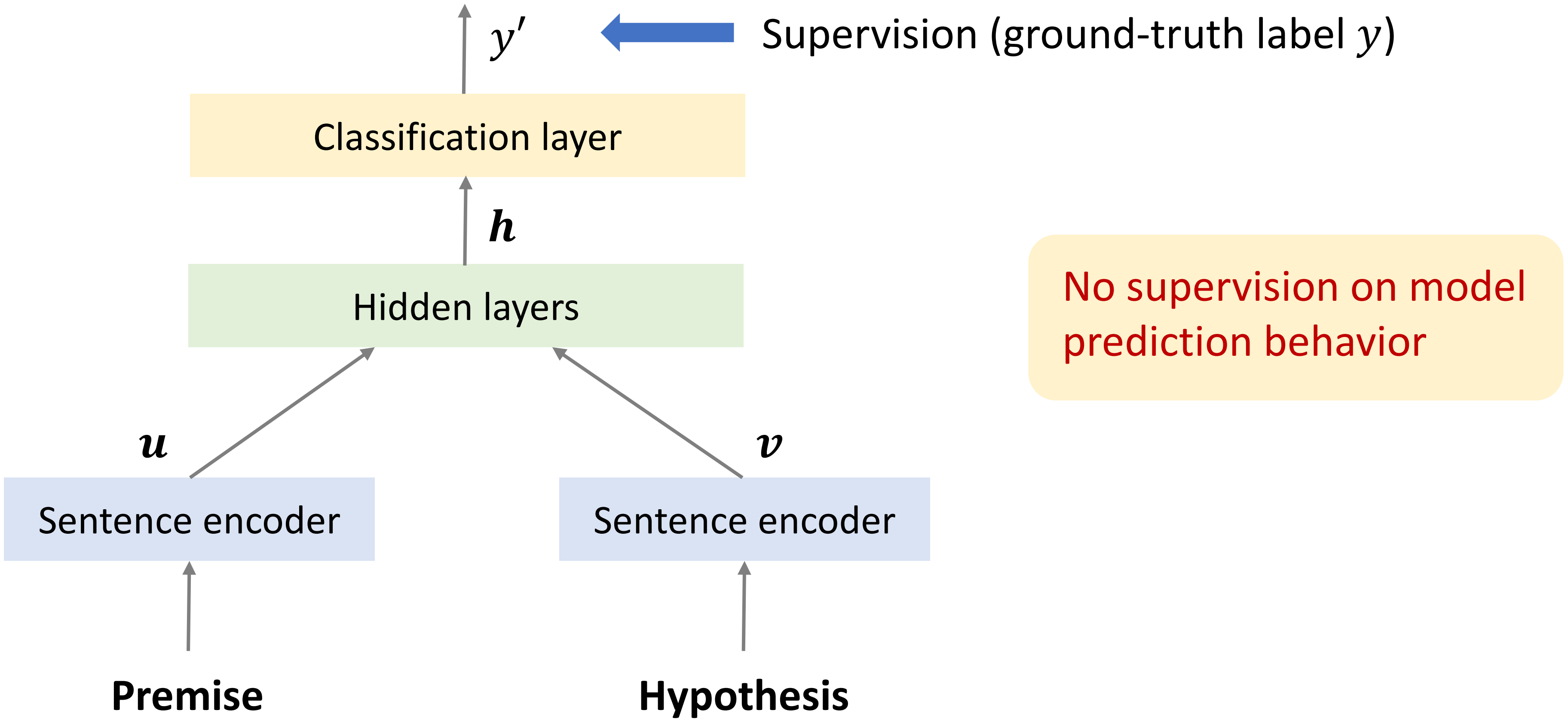
e-SNLI

- **Model training on NLI**



e-SNLI

- **Model training on NLI**



e-SNLI

- **Undesirable prediction behavior**

Models use the lexical overlap between sentence pairs to blindly predict “entailment”

[McCoy et al., 2019]

Premise	Hypothesis	Label
The judge was paid by the actor	The actor paid the judge	entailment
The doctor near the actor danced	The doctor danced	entailment
The lawyer was advised by the actor	The actor advised the lawyer	entailment
The banker near the judge saw the actor	The banker saw the actor	entailment
⋮	⋮	⋮

e-SNLI

- **Undesirable prediction behavior**

Models use the lexical overlap between sentence pairs to blindly predict “entailment”

[McCoy et al., 2019]

Premise	Hypothesis	Label
The judge was paid by the actor	The actor paid the judge	entailment
The doctor near the actor danced	The doctor danced	entailment
The lawyer was advised by the actor	The actor advised the lawyer	entailment
The banker near the judge saw the actor	The banker saw the actor	entailment
⋮	⋮	⋮

Over 90% of the data support this heuristic

e-SNLI

- **Undesirable prediction behavior**

Models use the lexical overlap between sentence pairs to blindly predict “entailment”

[McCoy et al., 2019]

Premise	Hypothesis	Label
The judge was paid by the actor	The actor paid the judge	entailment
The doctor near the actor danced	The doctor danced	
The lawyer was advised by the actor	The actor advised the lawye	
The banker near the judge saw the actor	The banker saw the actor	
⋮	⋮	

Model performance drops significantly on challenging datasets (e.g., HANS)

Example

Premise cat chased mouse

Hypothesis mouse chased cat

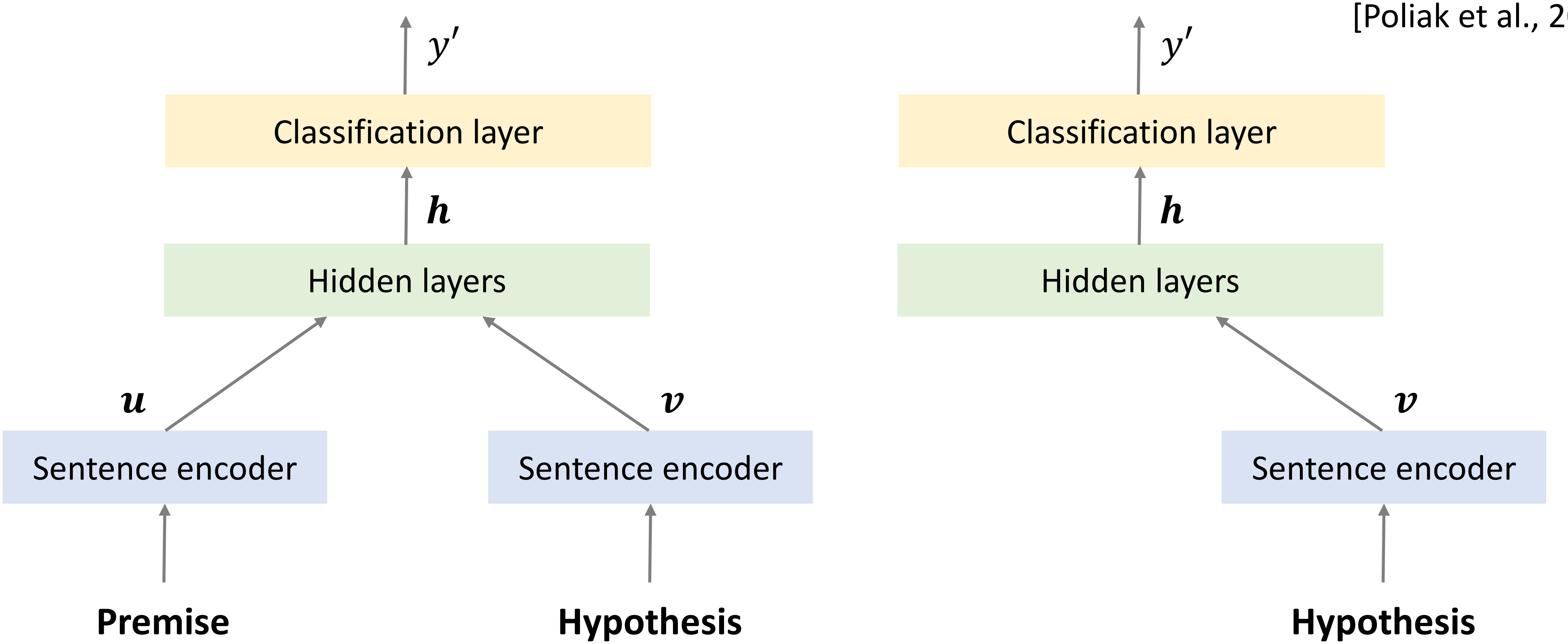
Prediction entailment

e-SNLI

- **Undesirable prediction behavior**

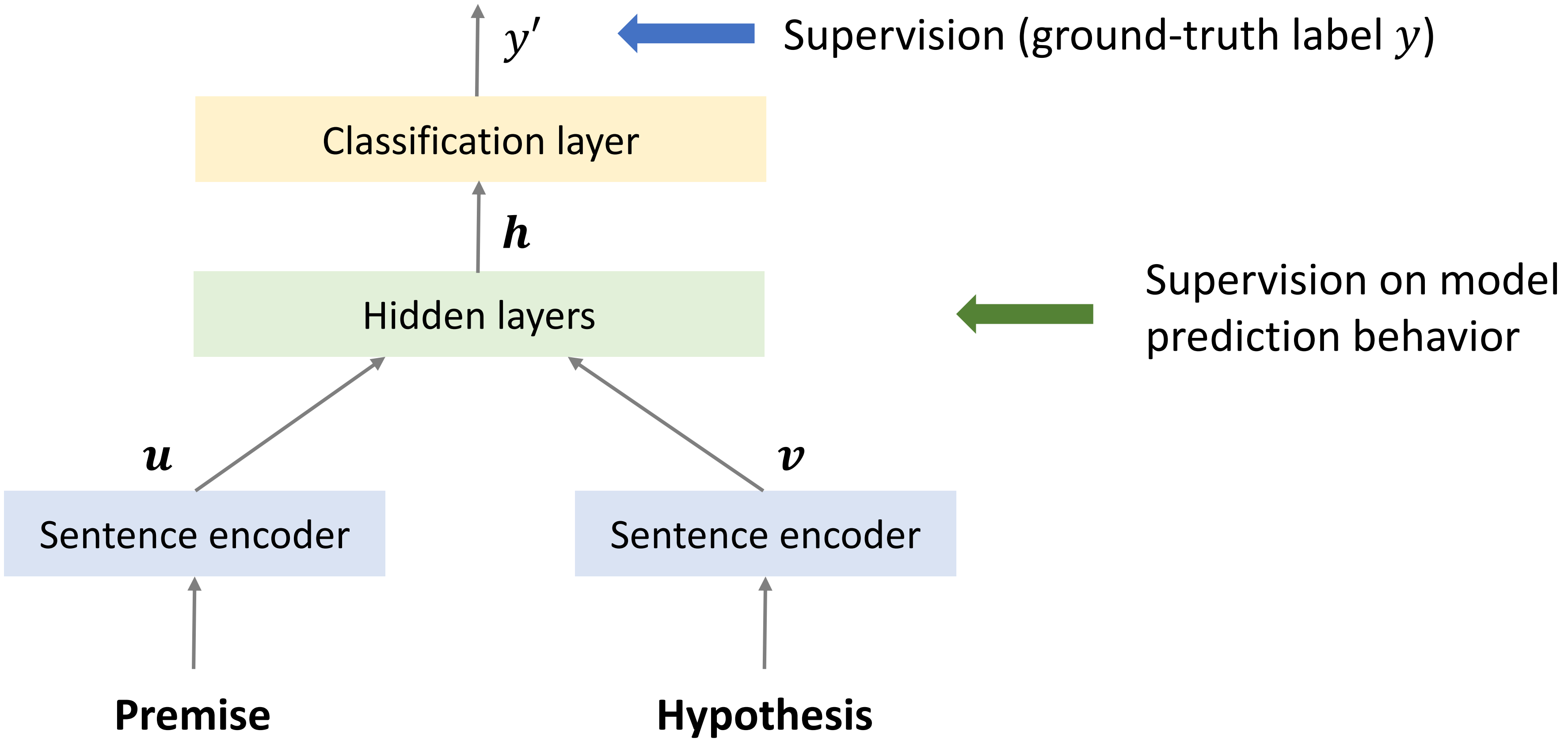
Models can achieve plausibly good performance by solely looking at **Hypothesis**

[Poliak et al., 2018]



e-SNLI

- **Model training on NLI with human-annotated explanations**



Question?

e-SNLI

Free-form natural language explanations

- Natural language is readily comprehensible to an end-user
- It is easiest for humans to provide free-form language
- Natural language justifications might eventually be mined from existing large-scale free-form text

e-SNLI

Collecting human-annotated explanation (Amazon Mechanical Turk)

- Annotators were given the premise, hypothesis, and label
- They highlighted the words that they considered essential for the label
- They also provided the explanation

Premise: An adult dressed in black holds a stick.
Hypothesis: An adult is walking away, empty-handed.
Label: contradiction
Explanation: Holds a stick implies using hands so it is not empty-handed.

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.
Hypothesis: A young mother is playing with her daughter in a swing.
Label: neutral
Explanation: Child does not imply daughter and woman does not imply mother.

Premise: A man in an orange vest leans over a pickup truck.
Hypothesis: A man is touching a truck.
Label: entailment
Explanation: Man leans over a pickup truck implies that he is touching it.

e-SNLI

Collecting human-annotated explanation (Amazon Mechanical Turk)

Control the quality of free-form annotations

- Encourage annotators to focus on the non-obvious elements that induce the given relation
- Entailment: require justifications of all the parts of the hypothesis that do not appear in the premise
- Neutral/Contradiction: consider an explanation correct, if at least one element stated contributes to the relation
- Provide self-contained explanations

Example

"Anyone can knit, not just women." ✓

"It cannot be inferred they are women." ✗

e-SNLI

Collecting human-annotated explanation (Amazon Mechanical Turk)

Control the quality of free-form annotations

- Encourage annotators to focus on the non-obvious elements that induce the given relation
- Entailment: require justifications of all the parts of the hypothesis that do not appear in the premise
- Neutral/Contradiction: consider an explanation correct, if at least one element stated contributes to the relation
- Provide self-contained explanations

Example

"Anyone can knit, not just women." ✓

"It cannot be inferred they are women." ✗

- ✓ Filter out incorrect annotations
- ✓ Analyze and refine the collected data

e-SNLI

Collecting human-annotated explanation (Amazon Mechanical Turk)

SNLI

Train	Dev	Test
500K	5K	5K

1 explanation → 1 training example

3 explanations → 1 dev/test example

6325 workers with an average of 860 explanations per worker and a standard deviation of 403

Experiments

e-SNLI

Premise: An adult dressed in black holds a stick. Hypothesis: An adult is walking away, empty-handed. Label: contradiction Explanation: Holds a stick implies using hands so it is not empty-handed.
Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her. Hypothesis: A young mother is playing with her daughter in a swing. Label: neutral Explanation: Child does not imply daughter and woman does not imply mother.
Premise: A man in an orange vest leans over a pickup truck. Hypothesis: A man is touching a truck. Label: entailment Explanation: Man leans over a pickup truck implies that he is touching it.

Experiments

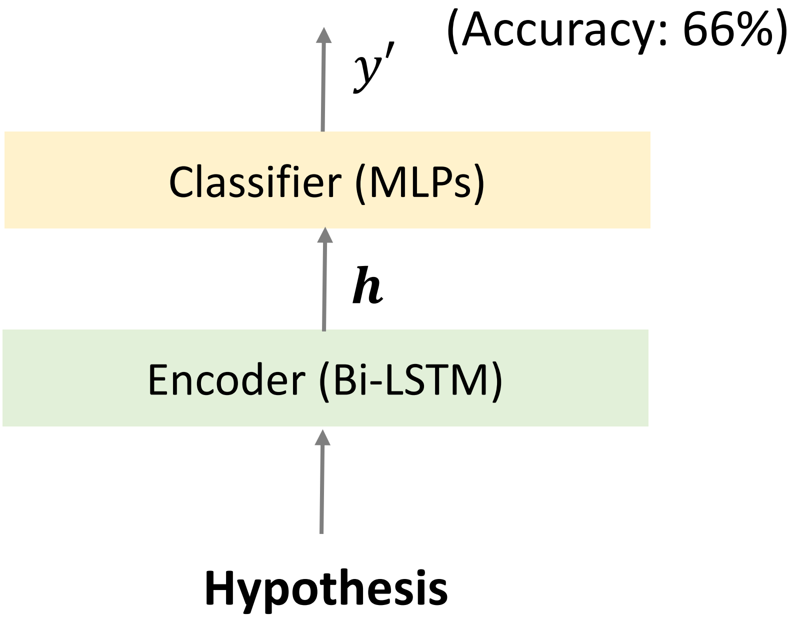
Research Questions

- PremiseAgnostic: a model that relies on artifacts to provide correct labels can provide correct explanations?
- PredictAndExplain: models trained on e-SNLI can predict a label and generate an explanation for the predicted label?
- ExplainThenPredict: models trained on e-SNLI can generate an explanation then predict the label given only the generated explanation?
- REPRESENT: models trained on e-SNLI can learn better universal sentence representations?
- TRANSFER: models trained on e-SNLI can transfer to out-of-domain NLI datasets?

Experiments

PremiseAgnostic: a model that relies on artifacts to provide correct labels can provide correct explanations?

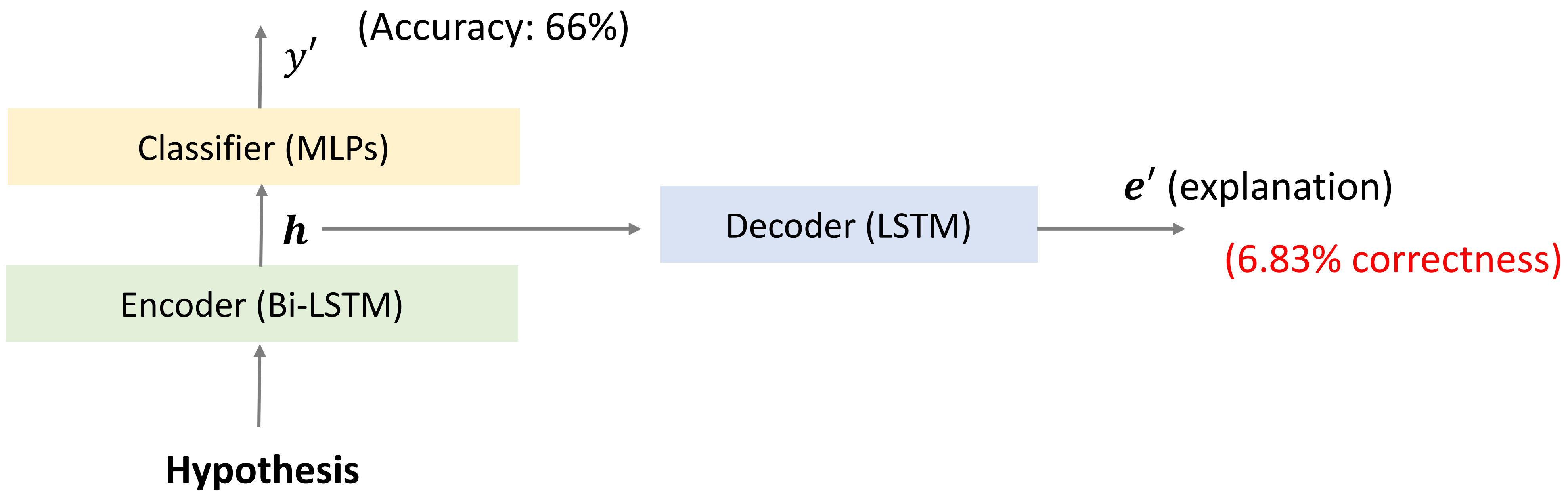
Models can achieve plausibly good performance by solely looking at **Hypothesis**



Experiments

PremiseAgnostic: a model that relies on artifacts to provide correct labels can provide correct explanations?

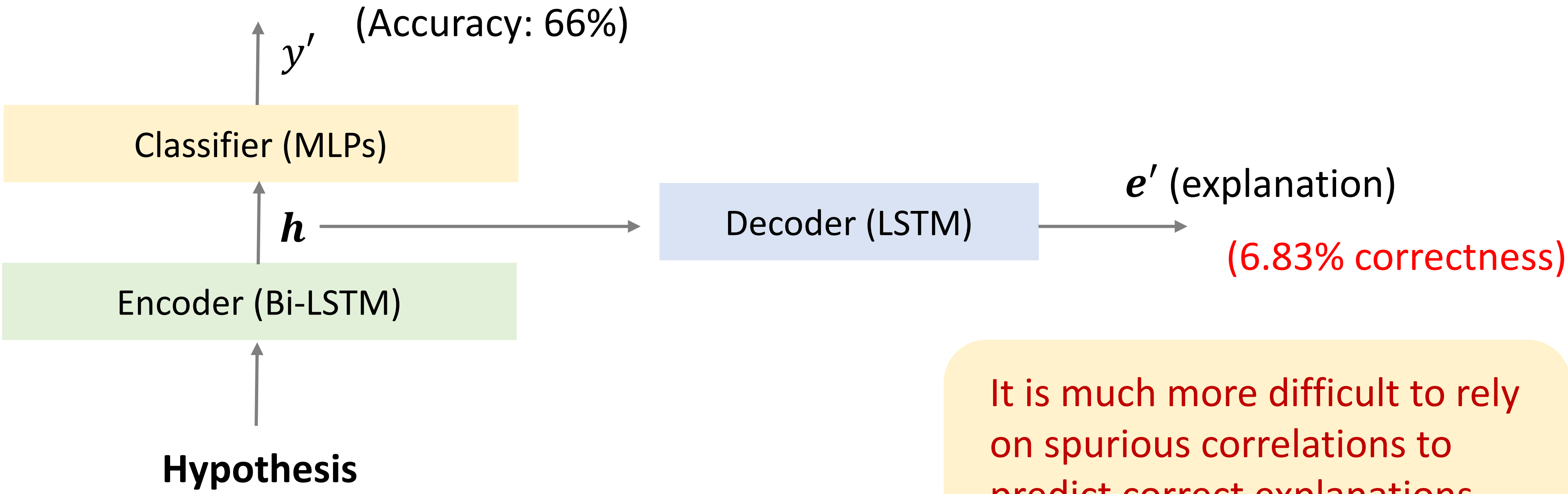
Models can achieve plausibly good performance by solely looking at **Hypothesis**



Experiments

PremiseAgnostic: a model that relies on artifacts to provide correct labels can provide correct explanations?

Models can achieve plausibly good performance by solely looking at **Hypothesis**



It is much more difficult to rely on spurious correlations to predict correct explanations than to predict correct labels

Experiments

Research Questions

- PremiseAgnostic: a model that relies on artifacts to provide correct labels can provide correct explanations?

- PredictAndExplain: models trained on e-SNLI can predict a label and generate an explanation for the predicted label?

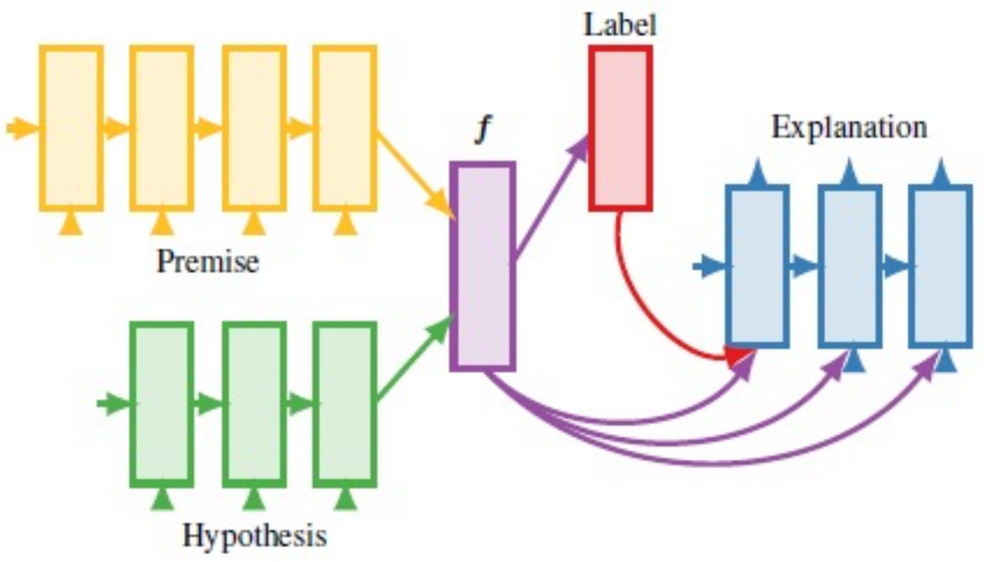
- ExplainThenPredict: models trained on e-SNLI can generate an explanation then predict the label given only the generated explanation?

- REPRESENT: models trained on e-SNLI can learn better universal sentence representations?

- TRANSFER: models trained on e-SNLI can transfer to out-of-domain NLI datasets?

Experiments

PredictAndExplain: models trained on e-SNLI can predict a label and generate an explanation for the predicted label?

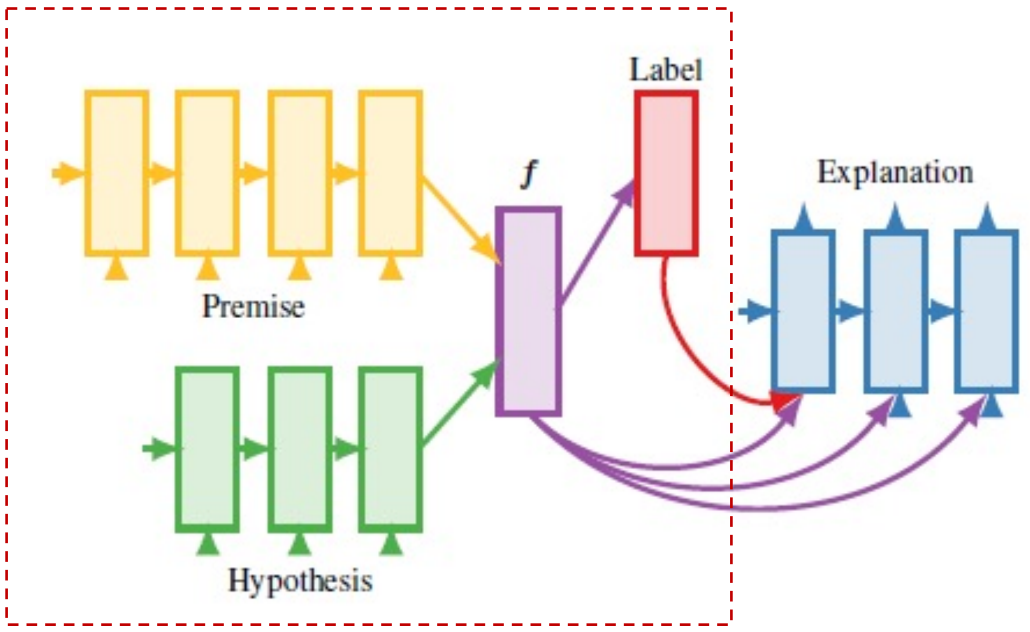


$$\mathcal{L} = \alpha \mathcal{L}_{label} + (1 - \alpha)\mathcal{L}_{explanation}$$

(negative log-likelihood)

Experiments

PredictAndExplain: models trained on e-SNLI can predict a label and generate an explanation for the predicted label?



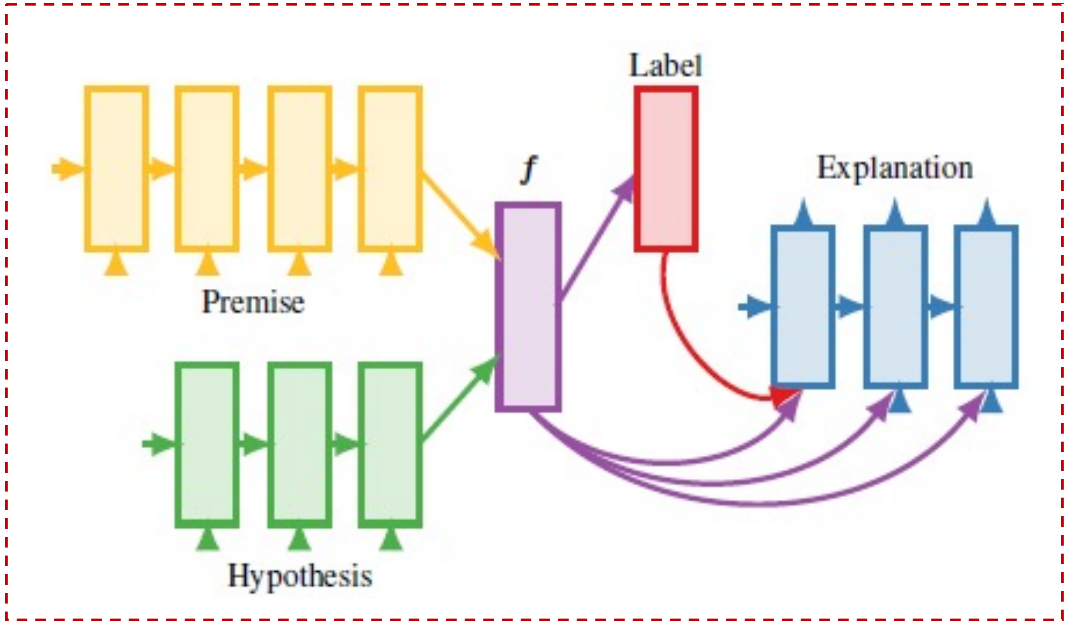
InferSent: accuracy=84.01%

$$\mathcal{L} = \alpha\mathcal{L}_{label} + (1-\alpha)\mathcal{L}_{explanation}$$

(negative log-likelihood)

Experiments

PredictAndExplain: models trained on e-SNLI can predict a label and generate an explanation for the predicted label?



e-InferSent: accuracy=83.96%

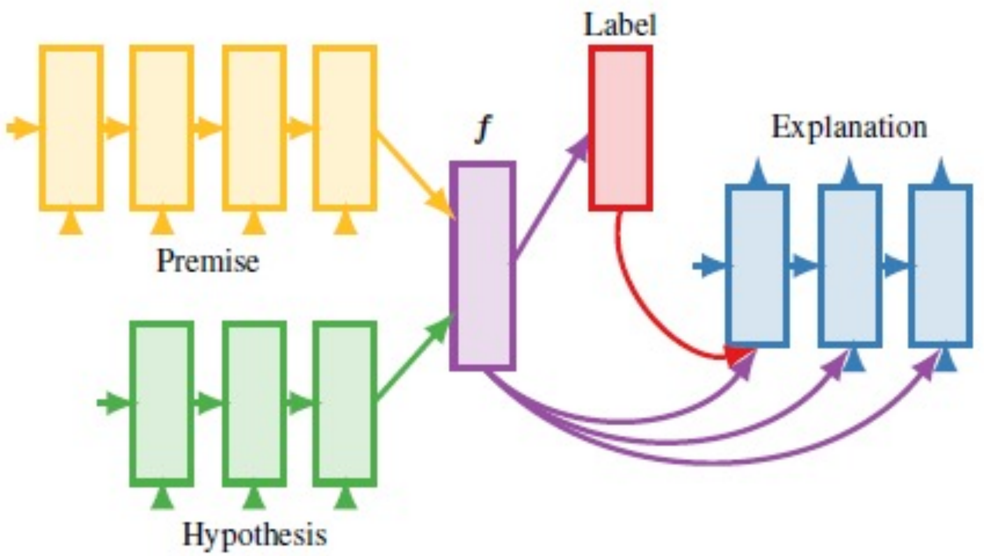
$$\mathcal{L} = \alpha \mathcal{L}_{label} + (1-\alpha)\mathcal{L}_{explanation}$$

(negative log-likelihood)

No sacrifice on label accuracy

Experiments

PredictAndExplain: models trained on e-SNLI can predict a label and generate an explanation for the predicted label?



$$\mathcal{L} = \alpha \mathcal{L}_{label} + (1 - \alpha)\mathcal{L}_{explanation}$$

(negative log-likelihood)

34.68% correct explanations

The best model was selected only based on the accuracy of the label classifier (not the perplexity of explanations)

Experiments

Research Questions

- PREMISEAGNOSTIC: a model that relies on artifacts to provide correct labels can provide correct explanations?
- PredictAndExplain: models trained on e-SNLI can predict a label and generate an explanation for the predicted label?
- ExplainThenPredict: models trained on e-SNLI can generate an explanation then predict the label given only the generated explanation?
- REPRESENT: models trained on e-SNLI can learn better universal sentence representations?
- TRANSFER: models trained on e-SNLI can transfer to out-of-domain NLI datasets?

Experiments

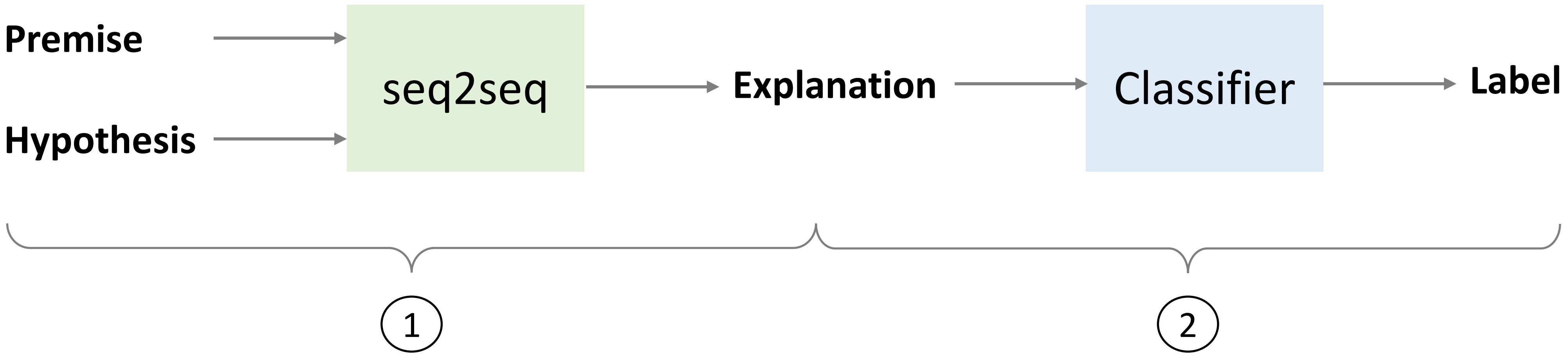
ExplainThenPredict: models trained on e-SNLI can generate an explanation then predict the label given only the generated explanation?

PredictAndExplain $p(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{y})$ → How the typical architecture used on SNLI can be adapted to justify its decisions in natural language

ExplainThenPredict $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{e})$ → Think of the explanation first and decide a label based on the explanation

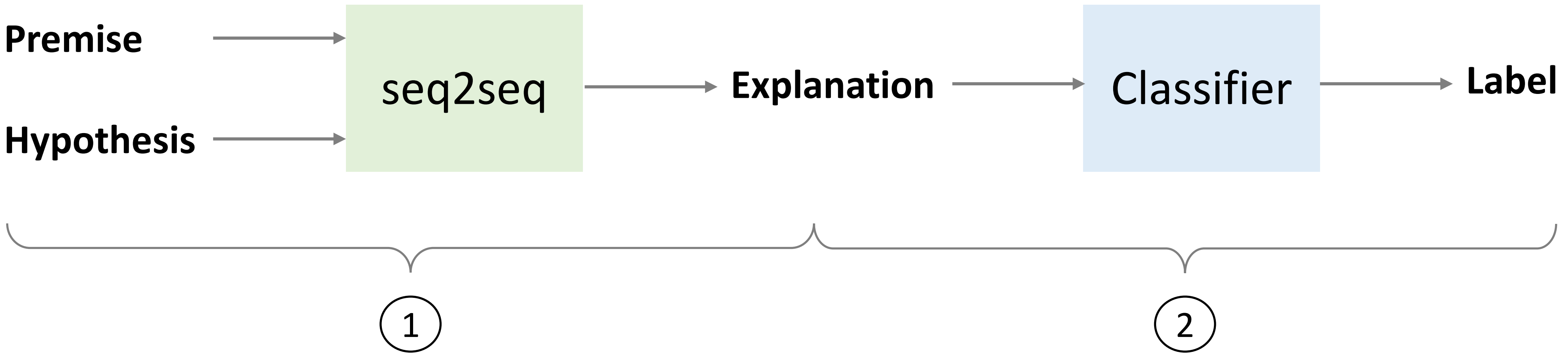
Experiments

ExplainThenPredict: models trained on e-SNLI can generate an explanation then predict the label given only the generated explanation?



Experiments

ExplainThenPredict: models trained on e-SNLI can generate an explanation then predict the label given only the generated explanation?

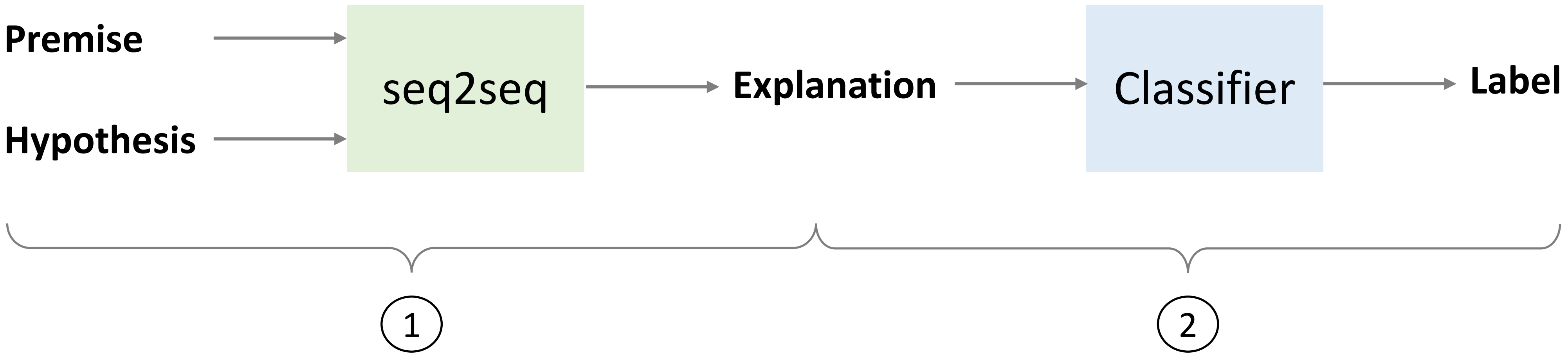


Test accuracy=81.59%

(drops from 83.96%)

Experiments

ExplainThenPredict: models trained on e-SNLI can generate an explanation then predict the label given only the generated explanation?

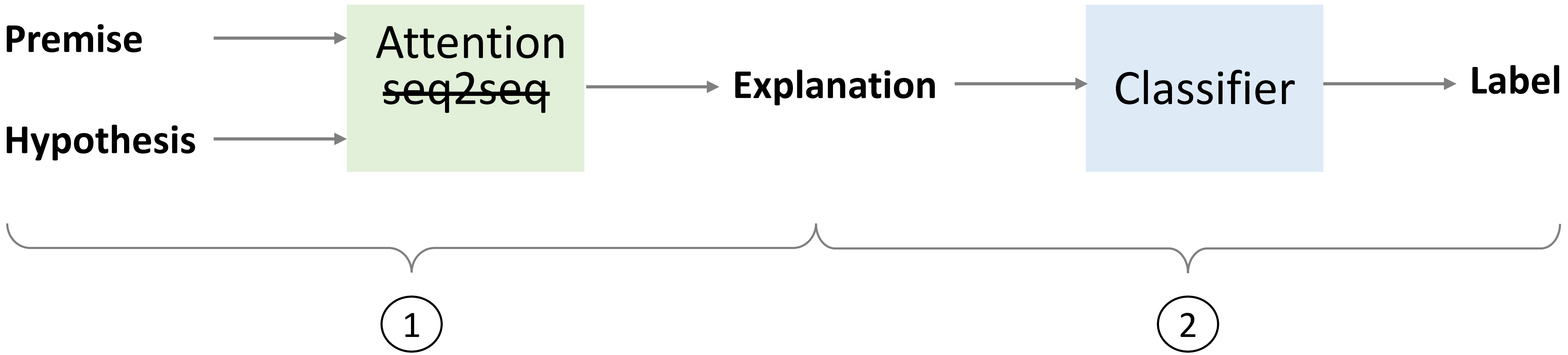


Correctness=49.8%
(increases from 34.68%)

Test accuracy=81.59%
(drops from 83.96%)

Experiments

ExplainThenPredict: models trained on e-SNLI can generate an explanation then predict the label given only the generated explanation?



Correctness=64.27%

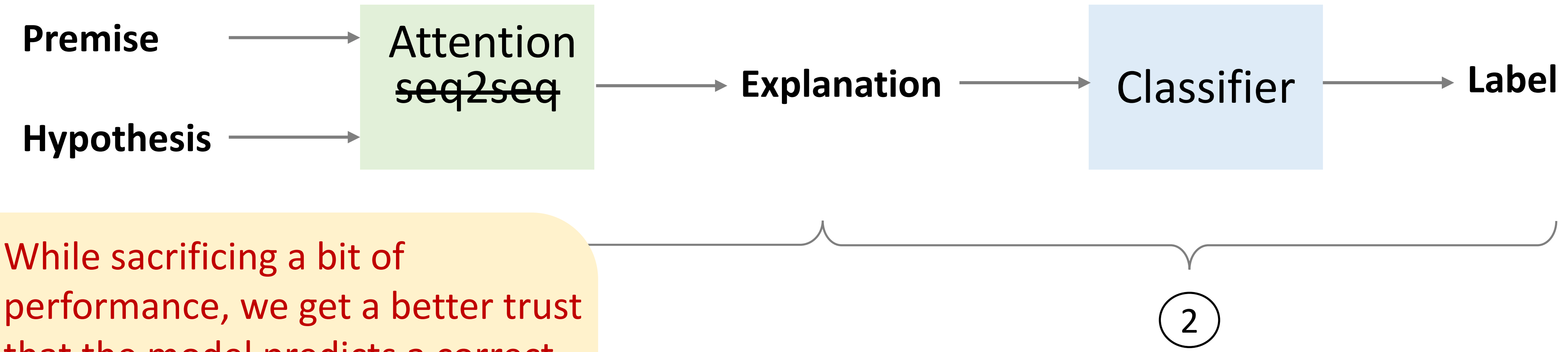
(increases from 34.68%)

Test accuracy=81.71%

(drops from 83.96%)

Experiments

ExplainThenPredict: models trained on e-SNLI can generate an explanation then predict the label given only the generated explanation?



- While sacrificing a bit of performance, we get a better trust that the model predicts a correct label based on the right reason
- It is still challenging to regularize a model prediction behavior even with human intervention

Correctness=64.27%
(increases from 34.68%)

Test accuracy=81.71%
(drops from 83.96%)

Experiments

Research Questions

- PREMISEAGNOSTIC: a model that relies on artifacts to provide correct labels can provide correct explanations?
- PredictAndExplain: models trained on e-SNLI can predict a label and generate an explanation for the predicted label?
- ExplainThenPredict: models trained on e-SNLI can generate an explanation then predict the label given only the generated explanation?
- REPRESENT: models trained on e-SNLI can learn better universal sentence representations?
- TRANSFER: models trained on e-SNLI can transfer to out-of-domain NLI datasets?

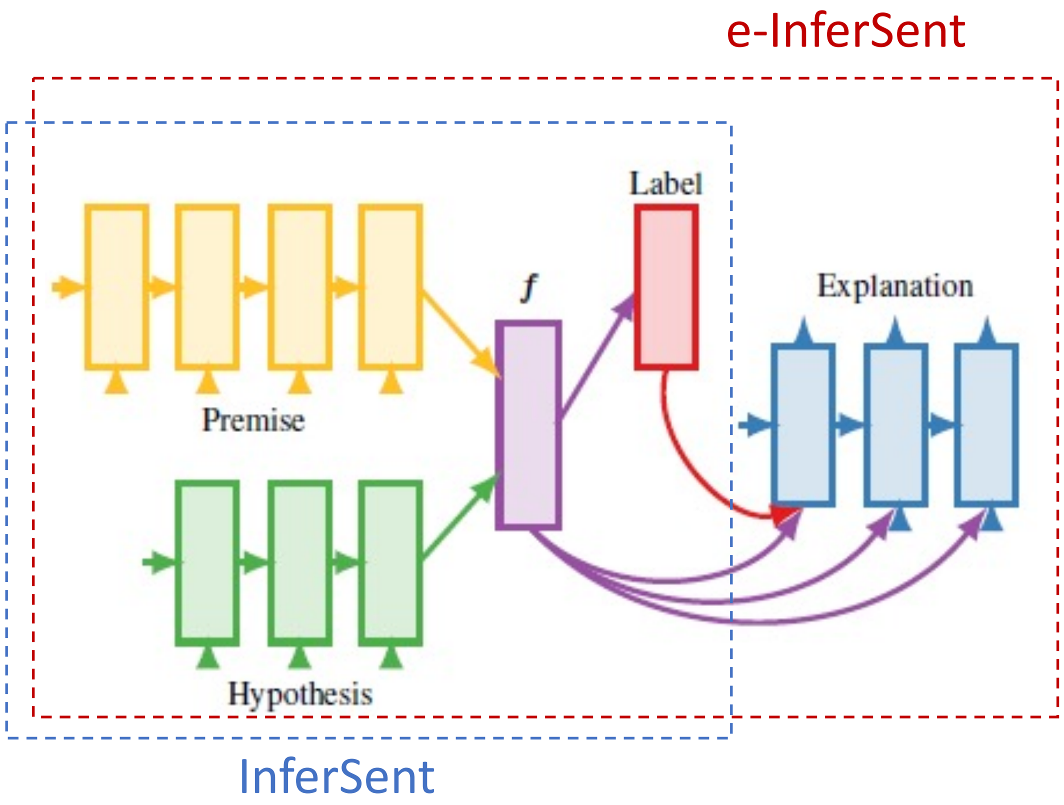
Experiments

REPRESENT: models trained on e-SNLI can learn better universal sentence representations?

Learning an encoder to provide semantically meaningful fixed-length representations of phrases/sentences

Experiments

REPRESENT: models trained on e-SNLI can learn better universal sentence representations?



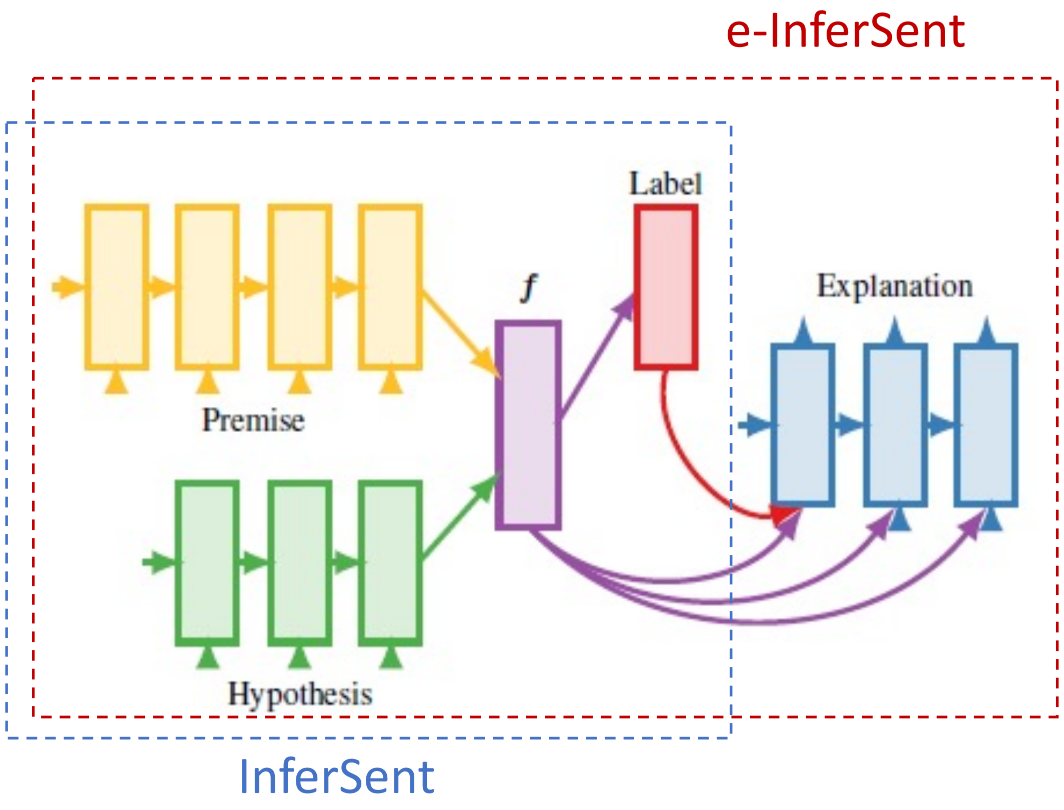
SentEval: 10 downstream tasks

[Conneau et al., 2017]

e-InferSent significantly outperforms InferSent on 4 tasks, while it is significantly outperformed only on 1 task

Experiments

REPRESENT: models trained on e-SNLI can learn better universal sentence representations?



SentEval: 10 downstream tasks

[Conneau et al., 2017]

e-InferSent significantly outperforms InferSent on 4 tasks, while it is significantly outperformed only on 1 task

Training with explanations helps the model to learn overall better sentence representations

Experiments

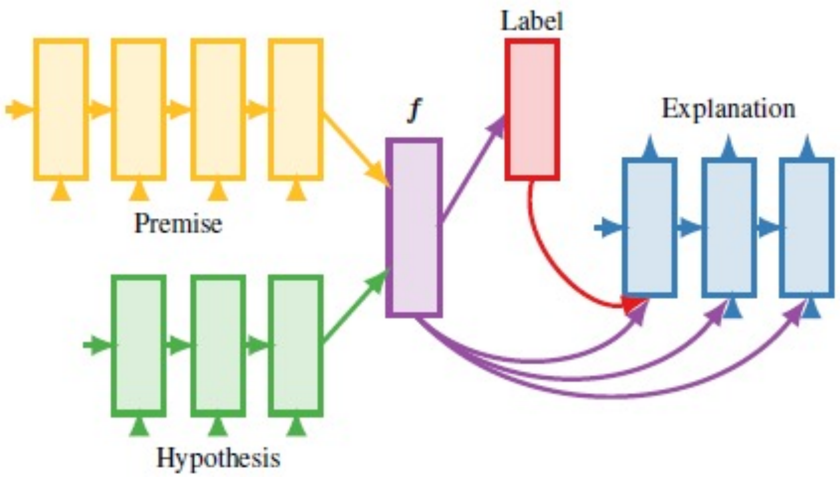
Research Questions

- PREMISEAGNOSTIC: a model that relies on artifacts to provide correct labels can provide correct explanations?
- PredictAndExplain: models trained on e-SNLI can predict a label and generate an explanation for the predicted label?
- ExplainThenPredict: models trained on e-SNLI can generate an explanation then predict the label given only the generated explanation?
- REPRESENT: models trained on e-SNLI can learn better universal sentence representations?
- TRANSFER: models trained on e-SNLI can transfer to out-of-domain NLI datasets?

Experiments

TRANSFER: models trained on e-SNLI can transfer to out-of-domain NLI datasets?

e-InferSent

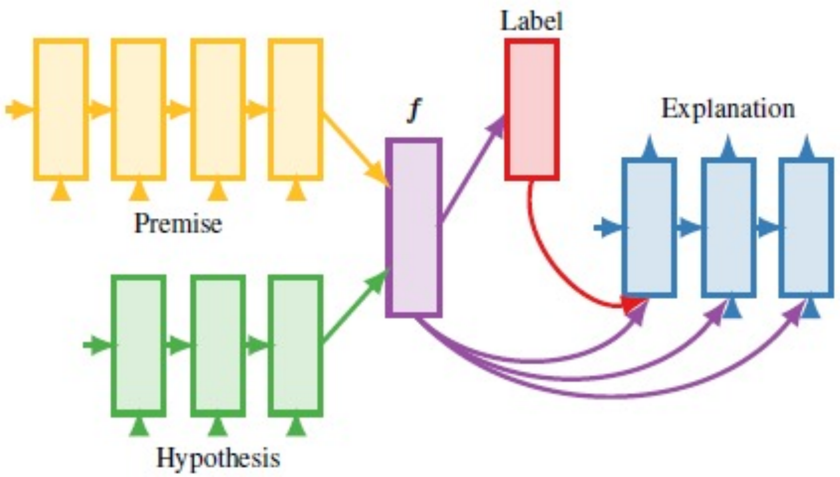


	SICK-E	MNLI
Test accuracy	53.54%	57.16%
Correct explanations	30.64%	1.92%

Experiments

TRANSFER: models trained on e-SNLI can transfer to out-of-domain NLI datasets?

e-InferSent



	SICK-E	MNLI
Test accuracy	53.54%	57.16%
Correct explanations	30.64%	1.92%

Transfer learning for generating explanations in out-of-domain NLI is still challenging

Discussion

- e-SNLI has been continuously studied in Explainable AI
- A good guideline: collecting human annotations, constructing a new dataset, comprehensive analyses…
- Collecting human annotated explanations is expensive
- Balancing model performance and interpretability

Improving Intrinsic Interpretability

- Training with rationales

- Variational word masks (VMASK)

VMASK

Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers

Hanjie Chen, Yangfeng Ji

(EMNLP, 2020)

Motivation

Models with similar network architectures have different interpretability

- Model A and B make the same and correct predictions
- LIME and SampleShapley generate different explanations for A and B

Model	Method	Text & Explanation	Prediction
A	LIME	An exceedingly clever piece of cinema	Positive
B		An exceedingly clever piece of cinema	Positive
A	SampleShapley	It becomes gimmicky instead of compelling	Negative
B		It becomes gimmicky instead of compelling	Negative

A
B

Motivation

Models with similar network architectures have different interpretability

- Model A and B make the same and correct predictions
- LIME and SampleShapley generate different explanations for A and B

Model	Method	Text & Explanation	Prediction
A	LIME	An exceedingly clever piece of cinema	Positive
B		An exceedingly clever piece of cinema	Positive
A	SampleShapley	It becomes gimmicky instead of compelling	Negative
B		It becomes gimmicky instead of compelling	Negative

A
B

Model B is more interpretable than A

Motivation

Models with similar network architectures have different interpretability

- Model A and B make the same and correct predictions
- LIME and SampleShapley generate different explanations for A and B

Model	Method	Text & Explanation	Prediction
A	LIME	An exceedingly clever piece of cinema	Positive
B		An exceedingly clever piece of cinema	Positive
A	SampleShapley	It becomes gimmicky instead of compelling	Negative
B		It becomes gimmicky instead of compelling	Negative

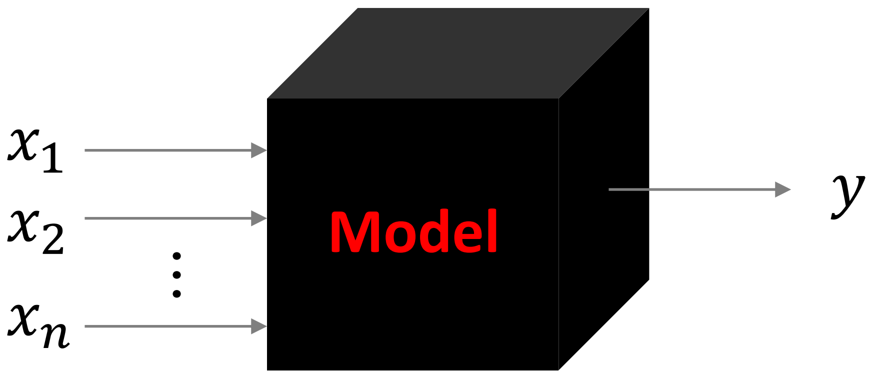
A
B

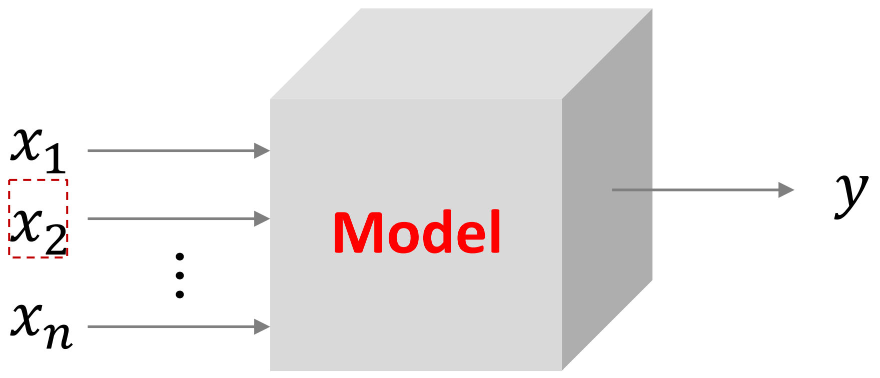
Model B is more interpretable than A

Improving the interpretability of existing models

How to improve model intrinsic interpretability?

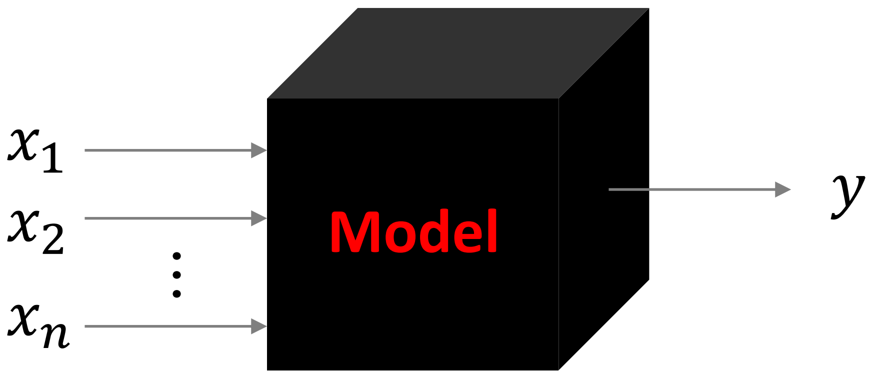
Teach the model to focus on important features to make predictions

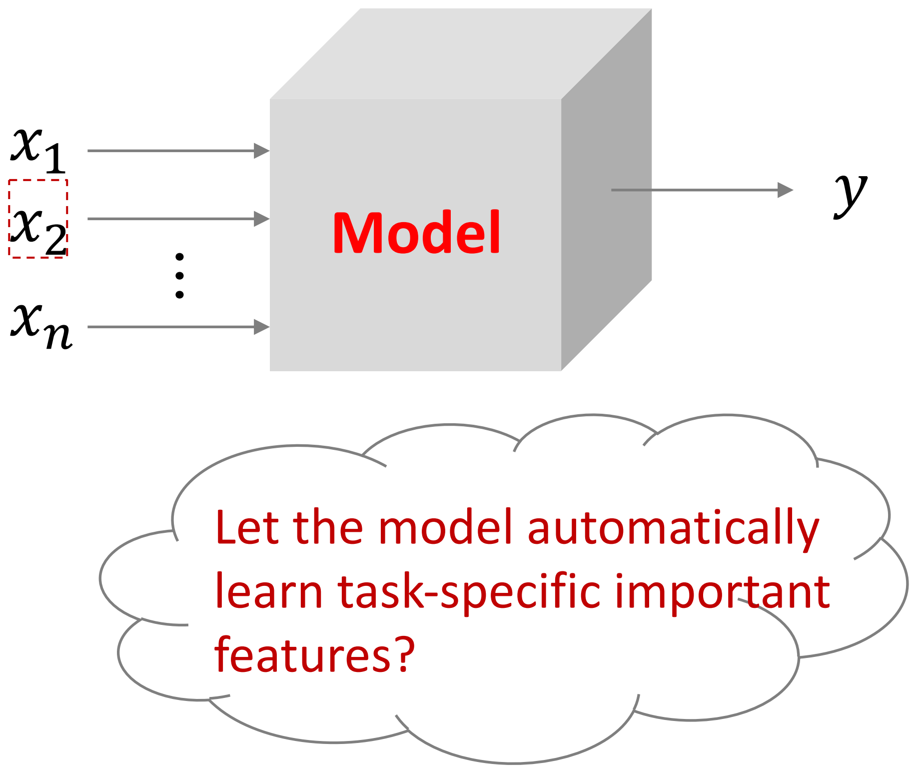




How to improve model intrinsic interpretability?

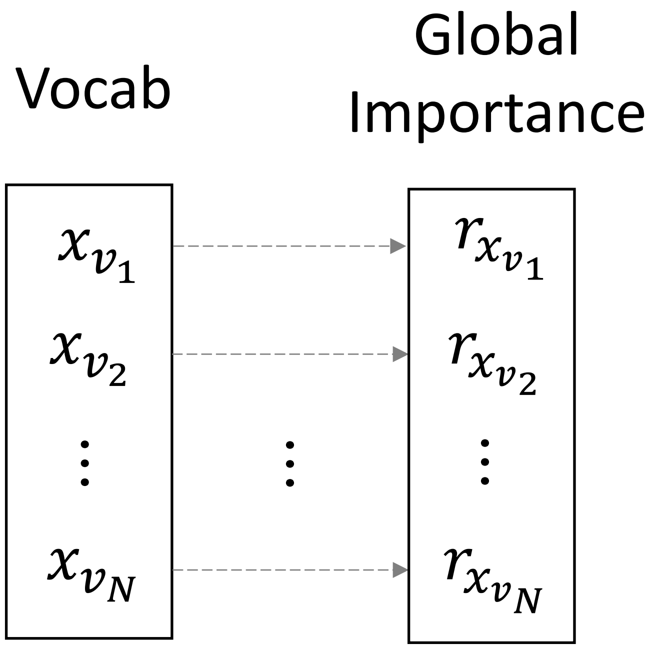
Teach the model to focus on important features to make predictions





VMASK

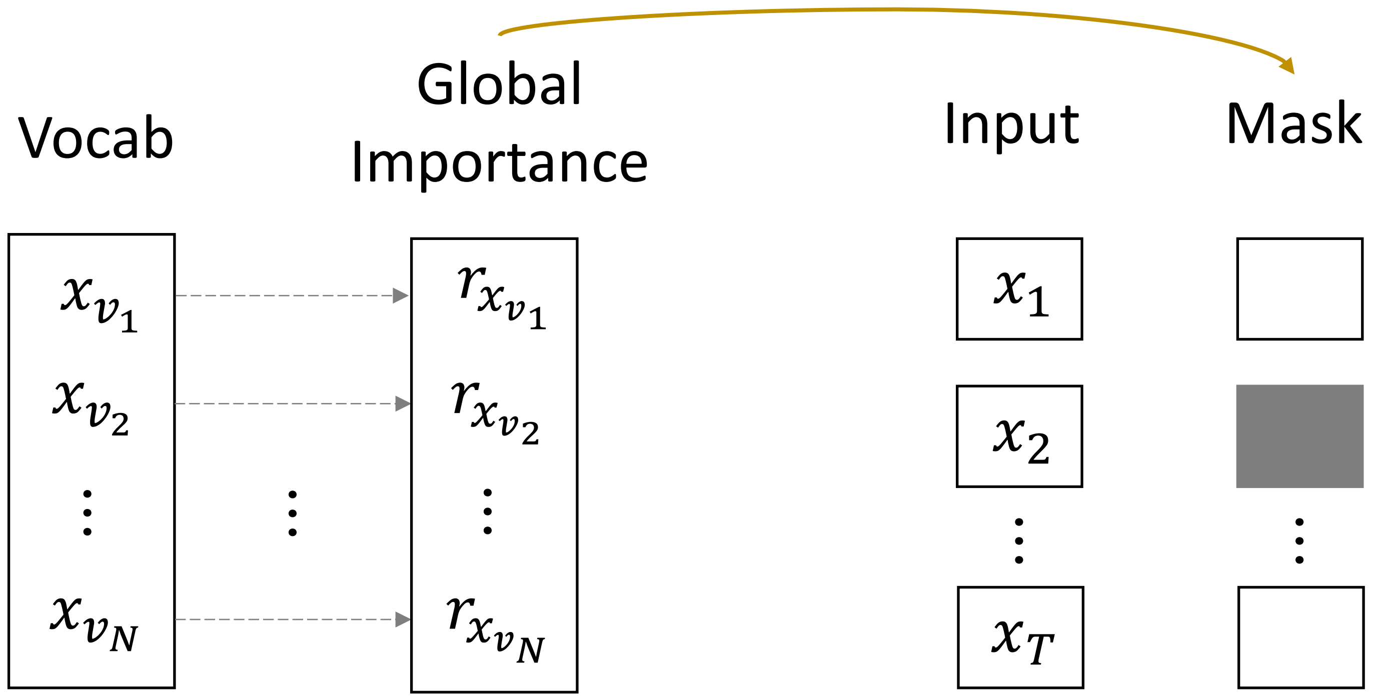
Teach the model to focus on important words to make predictions



$$r_{x_{v_i}} \in [0, 1]$$

VMASK

Teach the model to focus on important words to make predictions

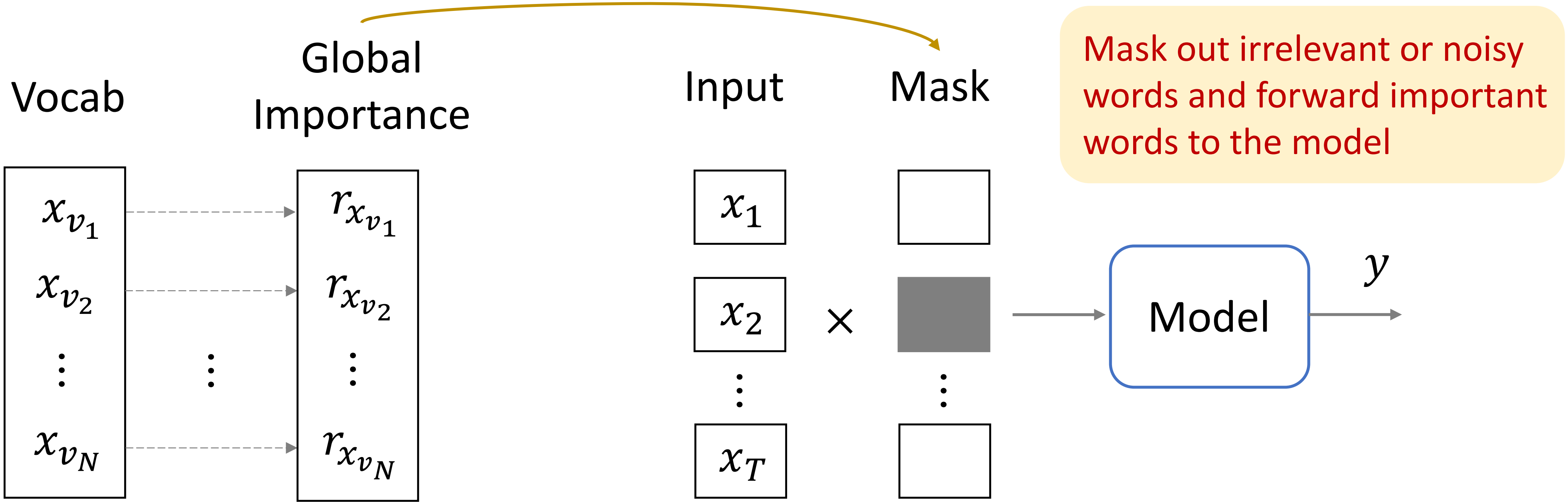


$r_{x_{v_i}} \in [0, 1]$

$R_{x_t} \in \{0, 1\}$ follows $Bernoulli(r_{x_t})$

VMASK

Teach the model to focus on important words to make predictions

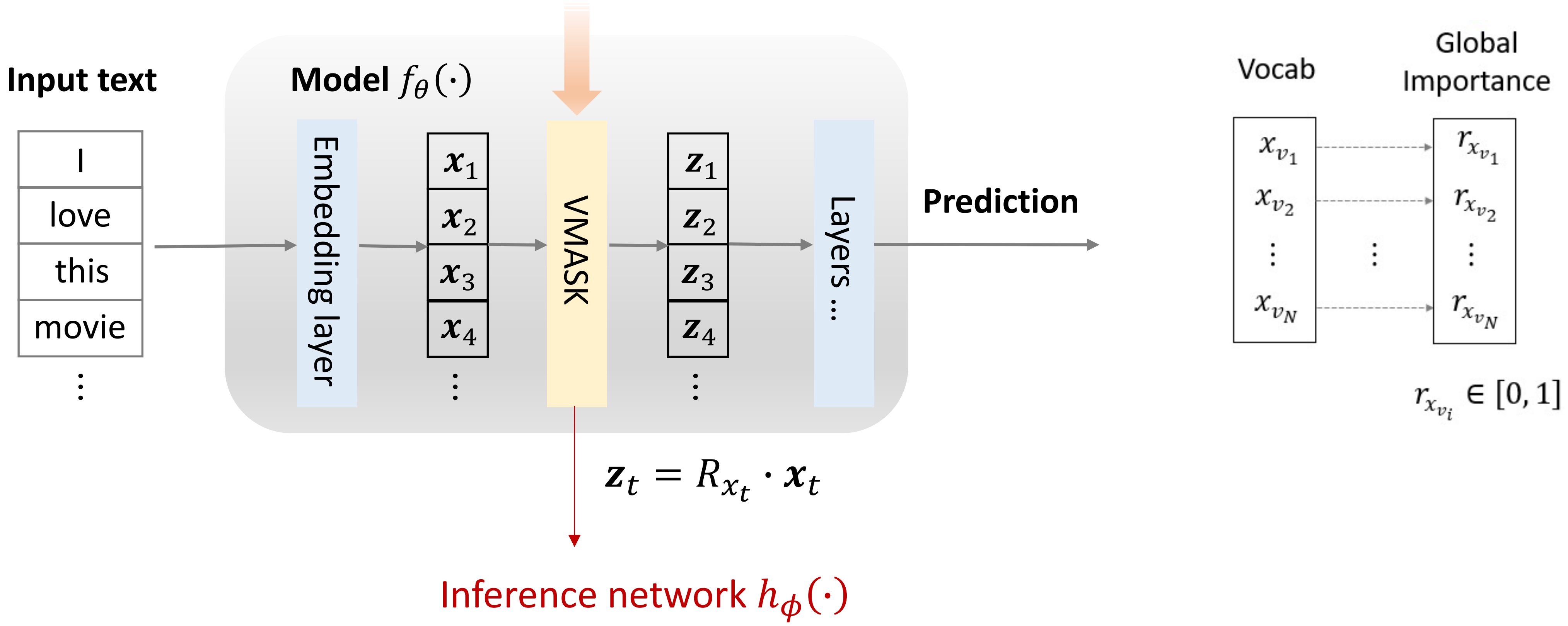


$r_{x_{v_i}} \in [0, 1]$

$R_{x_t} \in \{0, 1\}$ follows $Bernoulli(r_{x_t})$

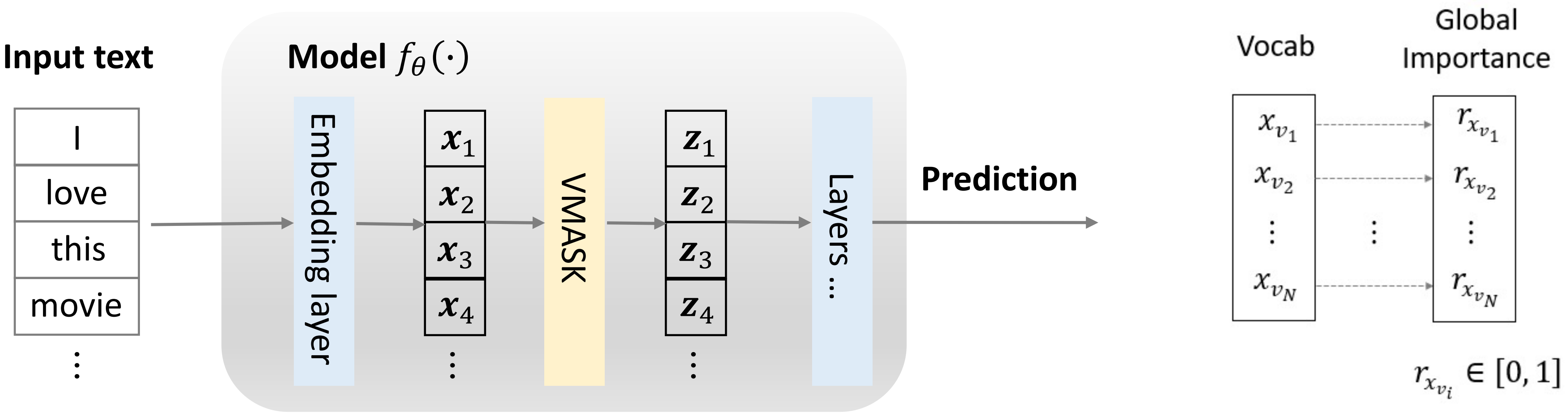
Variational Word Masks (VMASK)

Insert VMASK after word embedding layer and train it with the model jointly



Variational Word Masks (VMASK)

VMASK: remove redundant information from the input while keeping important information for model prediction



Information bottleneck

$$\max_{\boldsymbol{Z}} I(\boldsymbol{Z};\boldsymbol{Y}) - \beta \cdot I(\boldsymbol{Z};\boldsymbol{X})$$

$\boldsymbol{Z} = \boldsymbol{R} \odot \boldsymbol{x}$

$I(\cdot)$: mutual information

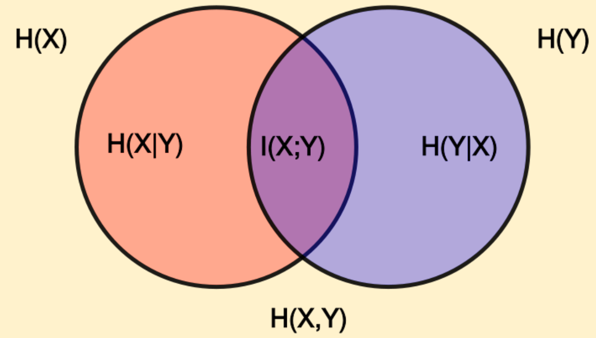
Information Theory

Information bottleneck [Tishby et al., 2000]

$$\max_{\boldsymbol{Z}} I(\boldsymbol{Z};\boldsymbol{Y}) - \beta \cdot I(\boldsymbol{Z};\boldsymbol{X})$$

Mutual information

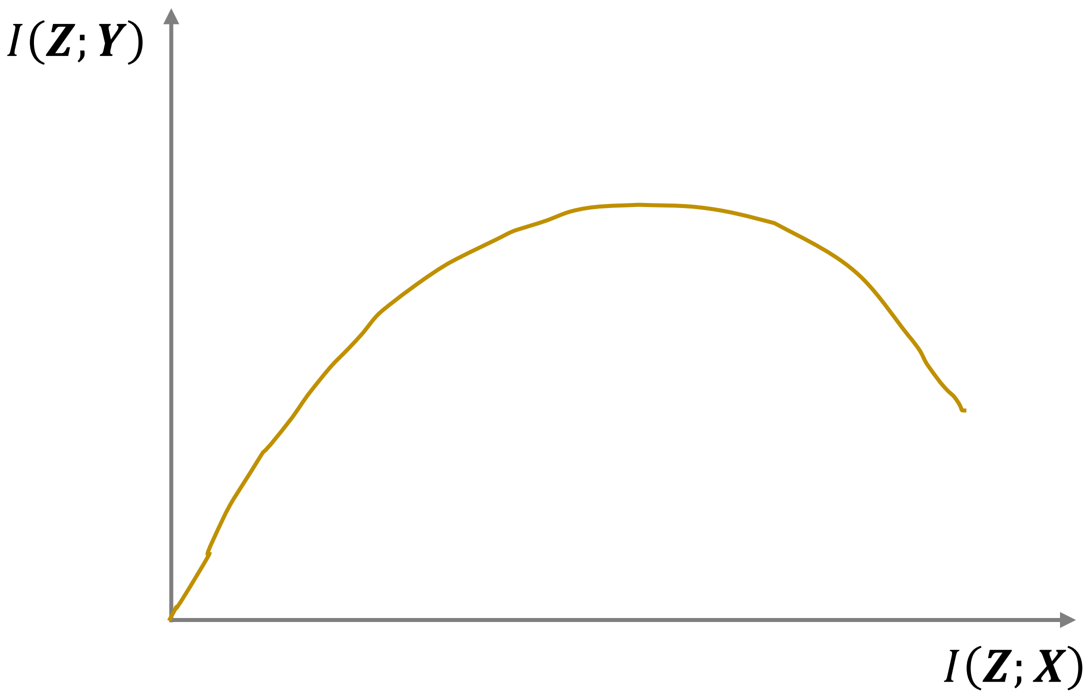
$$I(\boldsymbol{X};\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X})$$



Information Theory

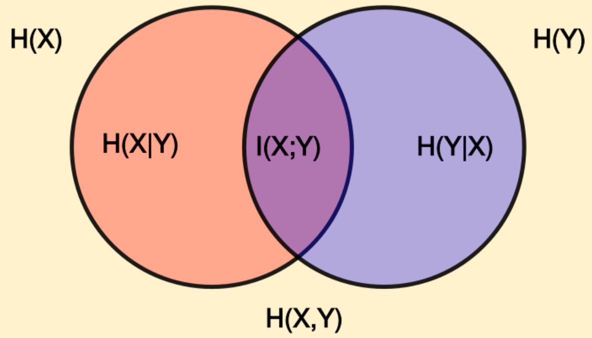
Information bottleneck [Tishby et al., 2000]

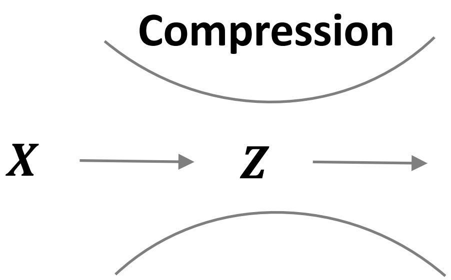
$$\max_{\boldsymbol{Z}} I(\boldsymbol{Z};\boldsymbol{Y}) - \beta \cdot I(\boldsymbol{Z};\boldsymbol{X})$$



Mutual information

$$I(\boldsymbol{X};\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X})$$





Information Theory

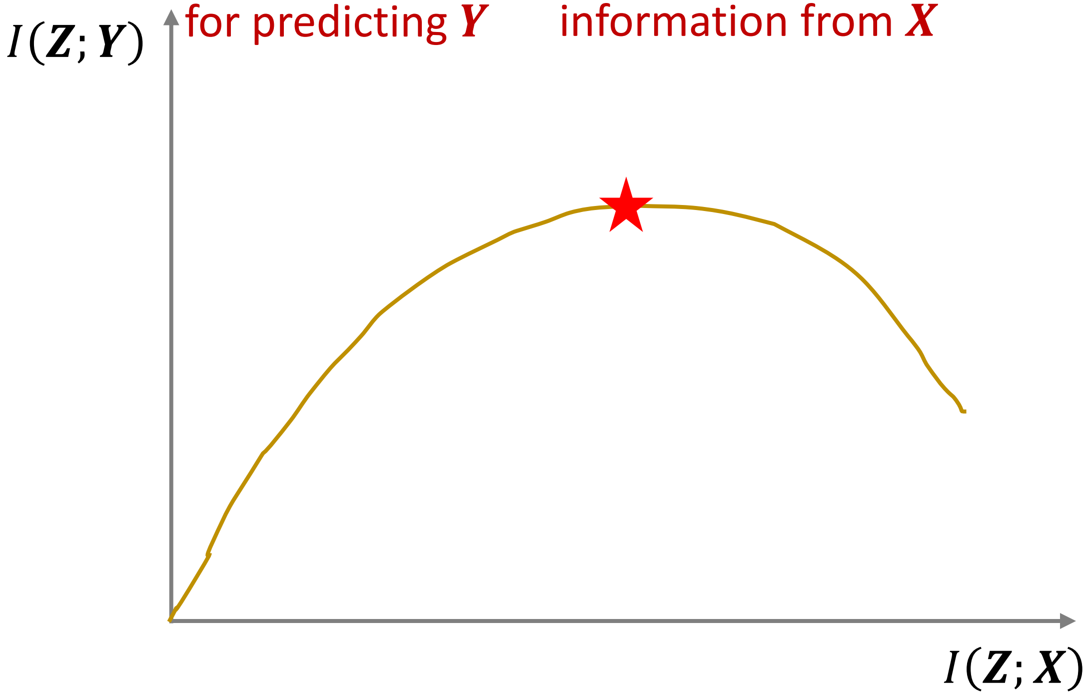
Information bottleneck

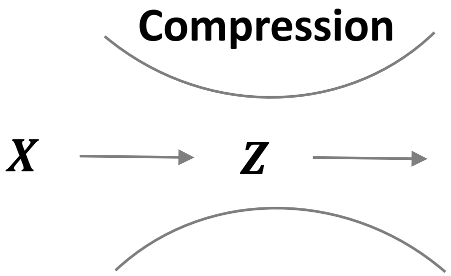
[Tishby et al., 2000]

$$\max_{\boldsymbol{Z}} \underline{I(\boldsymbol{Z};\boldsymbol{Y})} - \beta \cdot \underline{I(\boldsymbol{Z};\boldsymbol{X})}$$

Keep information for predicting $\boldsymbol{Y}$

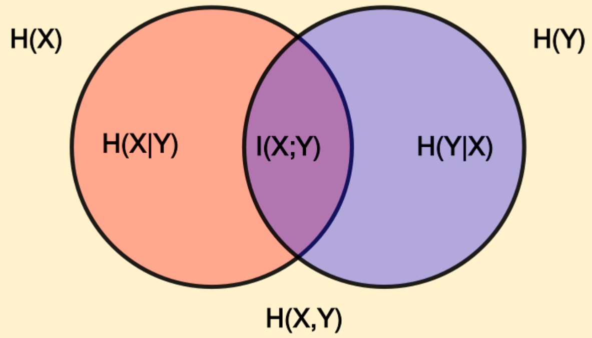
Remove redundant information from $\boldsymbol{X}$





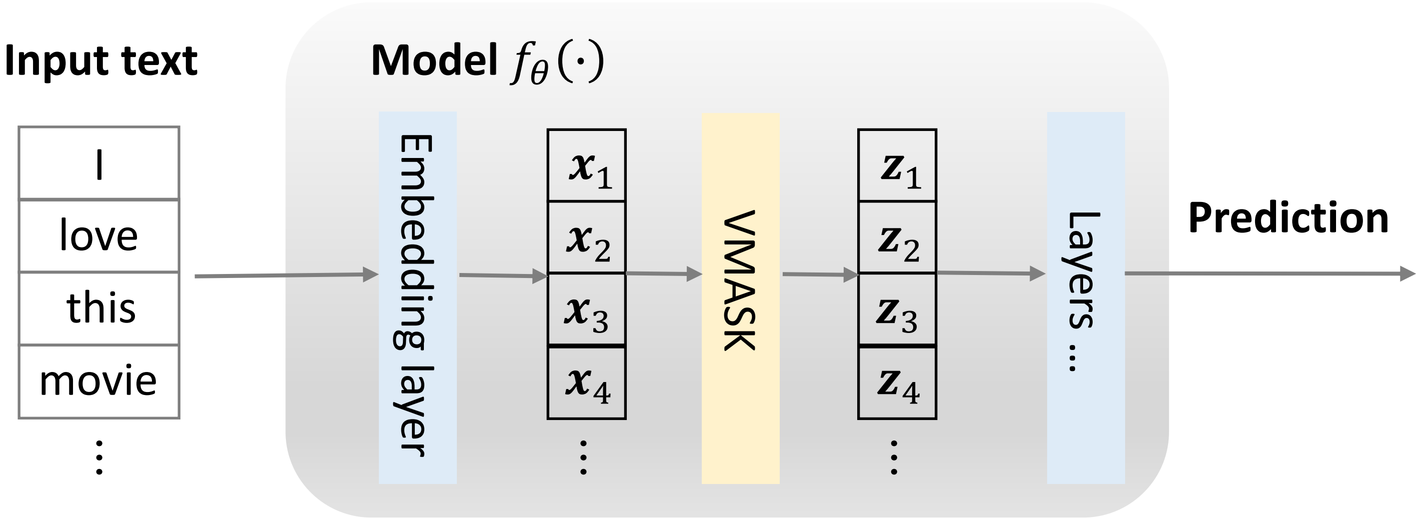
Mutual information

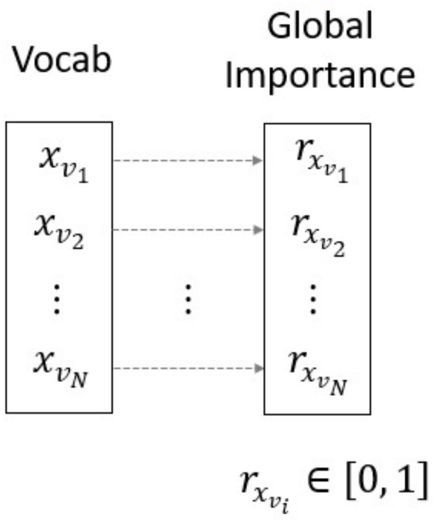
$$I(\boldsymbol{X};\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X})$$



Variational Word Masks (VMASK)

VMASK: remove redundant information from the input while keeping important information for model prediction





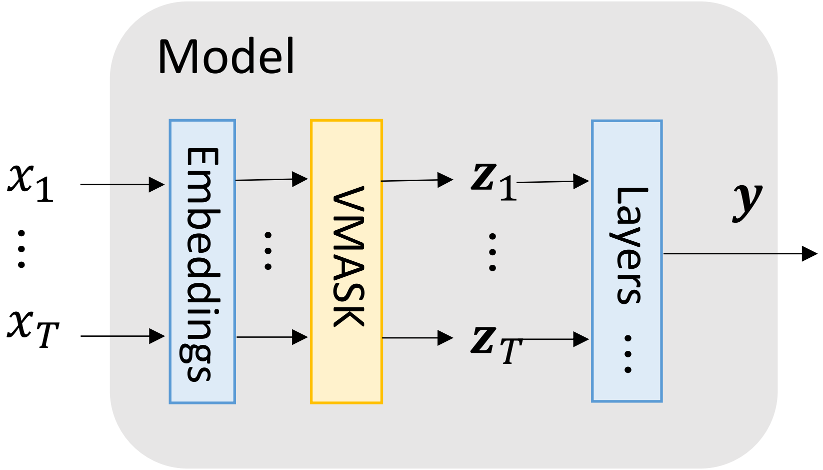
Information bottleneck

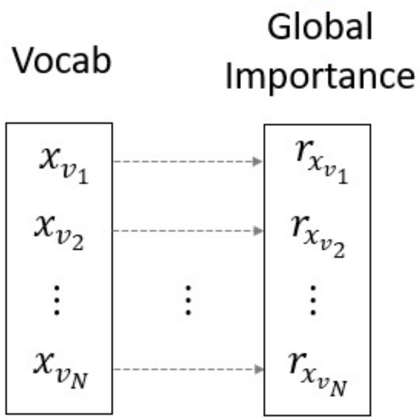
$$\max_{\boldsymbol{Z}} I(\boldsymbol{Z};\boldsymbol{Y}) - \beta \cdot I(\boldsymbol{Z};\boldsymbol{X})$$

$\boldsymbol{Z} = \boldsymbol{R} \odot \boldsymbol{x}$
$I(\cdot)$: mutual information

Optimizing IB makes the model prediction behavior more interpretable

Variational Word Masks (VMASK)





$r_{x_{v_i}} \in [0, 1]$

Information bottleneck

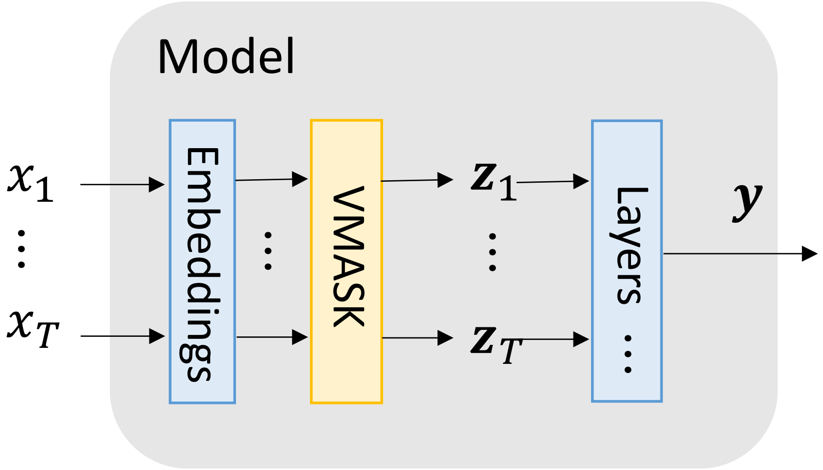
$\boldsymbol{Z} = \boldsymbol{R} \odot \boldsymbol{x}$
$I(\cdot)$: mutual information

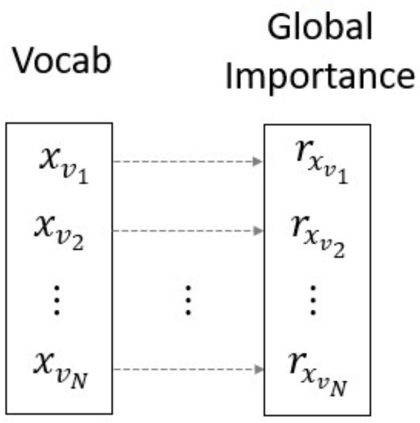
$$\max_{\boldsymbol{Z}} I(\boldsymbol{Z}; \boldsymbol{Y}) - \beta \cdot I(\boldsymbol{Z}; \boldsymbol{X})$$

Lower bound

$$\max_{\theta,\phi} \mathbb{E}_q \left[\log p\left(\boldsymbol{y}^{(i)} \middle| \boldsymbol{R}, \boldsymbol{x}^{(i)}\right)\right] + \beta \cdot H_q\left(\boldsymbol{R} \middle| \boldsymbol{x}^{(i)}\right)$$

Variational Word Masks (VMASK)





$r_{x_{v_i}} \in [0, 1]$

Information bottleneck

$$\max_{\boldsymbol{Z}} I(\boldsymbol{Z};\boldsymbol{Y}) - \beta \cdot I(\boldsymbol{Z};\boldsymbol{X})$$

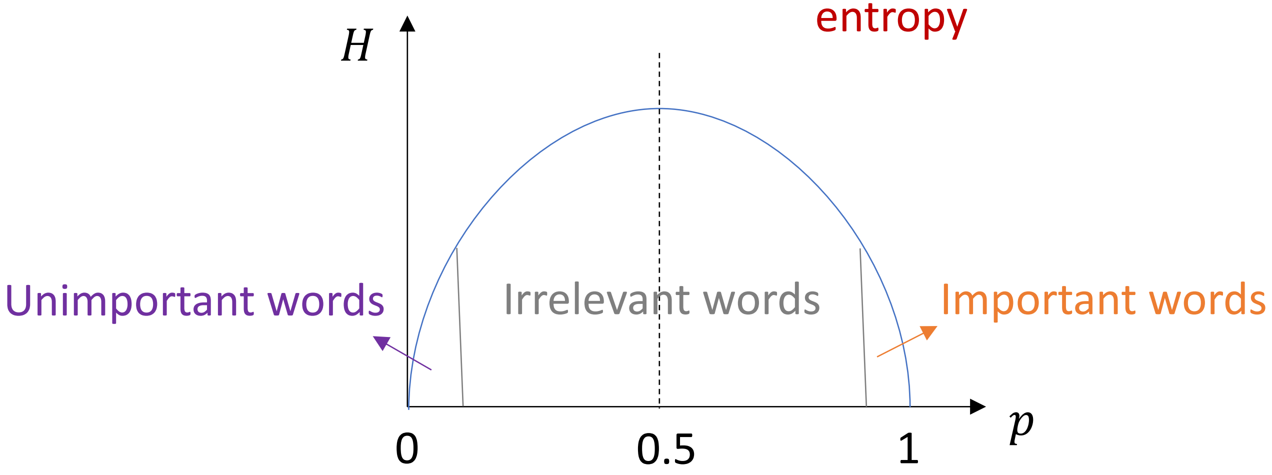
$\boldsymbol{Z} = \boldsymbol{R} \odot \boldsymbol{x}$

$I(\cdot)$: mutual information

Lower bound

$$\max_{\theta,\phi} \mathbb{E}_q \left[\log p\left(\boldsymbol{y}^{(i)} \middle| \boldsymbol{R}, \boldsymbol{x}^{(i)}\right)\right] + \beta \cdot H_q\left(\boldsymbol{R} \middle| \boldsymbol{x}^{(i)}\right)$$

entropy



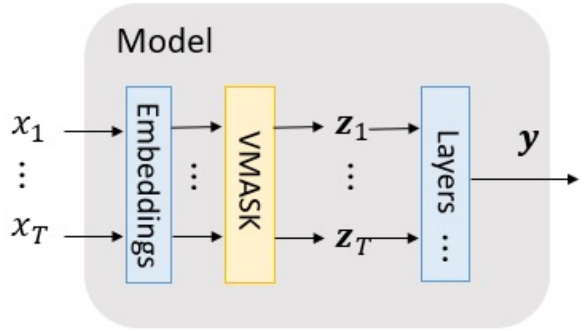
Summary

Goal

Improving interpretability: teaching the model to focus on important words to make predictions

VMASK

- Learn global word importance
- Generate binary word masks
- Mask out irrelevant or noisy words
- Keep important words for model prediction



Optimizing VMASK and model via Information Bottleneck

$$\max_{\boldsymbol{Z}} I(\boldsymbol{Z};\boldsymbol{Y}) - \beta \cdot I(\boldsymbol{Z};\boldsymbol{X})$$

$\boldsymbol{Z} = \boldsymbol{R} \odot \boldsymbol{x}$
$I(\cdot)$: mutual information

posterior $q_\phi\left(R_{x_{v_n}} \middle| \boldsymbol{x}_{v_n}\right) \longrightarrow \mathbb{E}[q_\phi(R_{x_t}|\boldsymbol{x}_t)]$ (global importance)

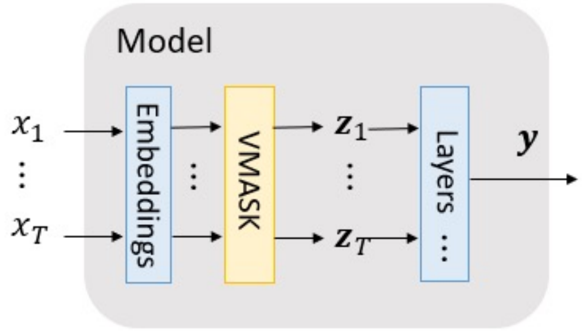
Summary

Goal

Improving interpretability: teaching the model to focus on important words to make predictions

VMASK

- Learn global word importance
- Generate binary word masks
- Mask out irrelevant or noisy words
- Keep important words for model prediction



Optimizing VMASK and model via Information Bottleneck

$$\max_{\boldsymbol{Z}} I(\boldsymbol{Z};\boldsymbol{Y}) - \beta \cdot I(\boldsymbol{Z};\boldsymbol{X})$$

$\boldsymbol{Z} = \boldsymbol{R} \odot \boldsymbol{x}$
$I(\cdot)$: mutual information

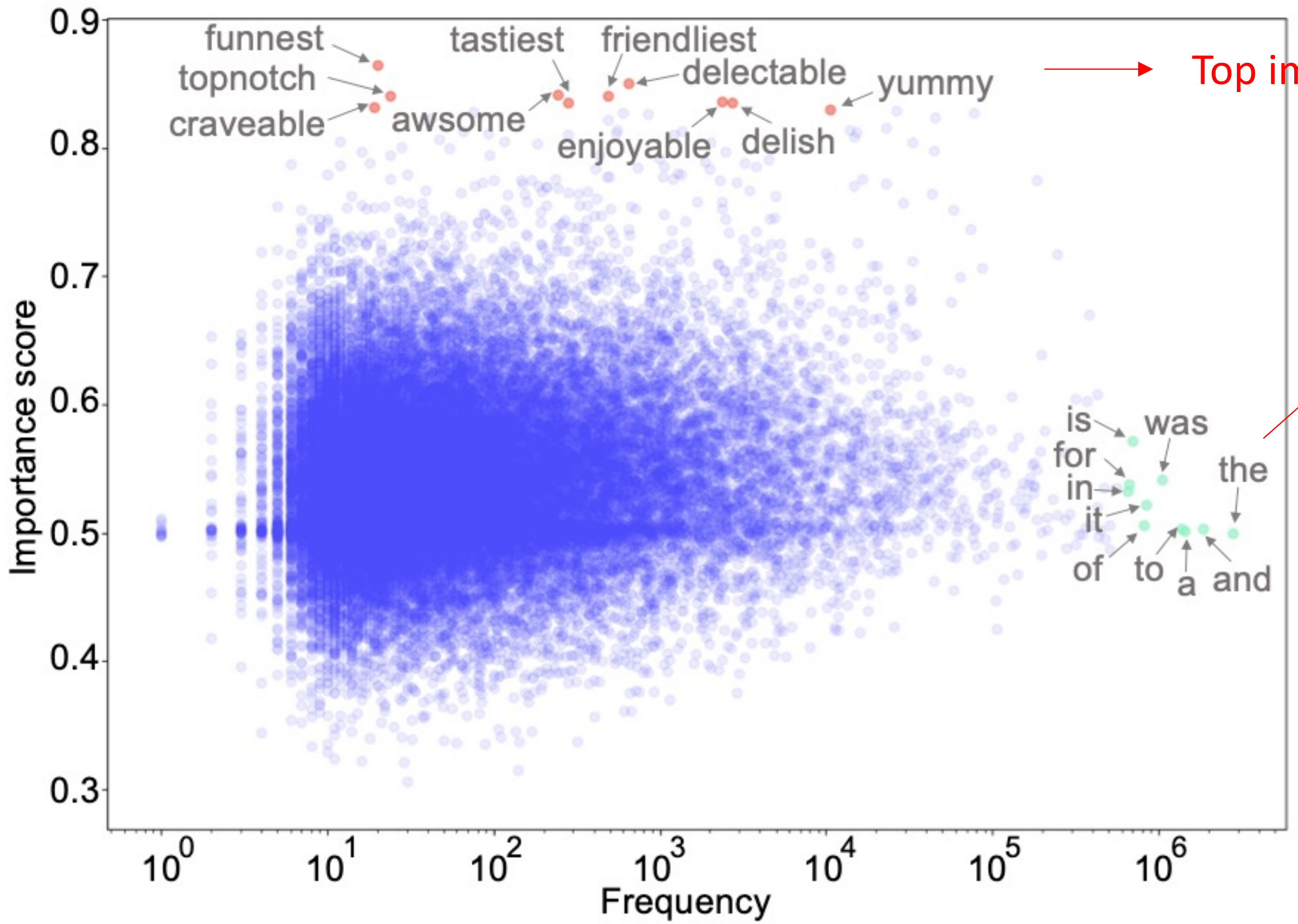
VMASK is model-agnostic

posterior $q_\phi\left(R_{x_{v_n}} \middle| \boldsymbol{x}_{v_n}\right) \longrightarrow \mathbb{E}[q_\phi(R_{x_t}|\boldsymbol{x}_t)]$ (global importance)

Question?

Global Word Importance vs. Frequency

- Sentiment classification: LSTM-VMASK on the Yelp dataset



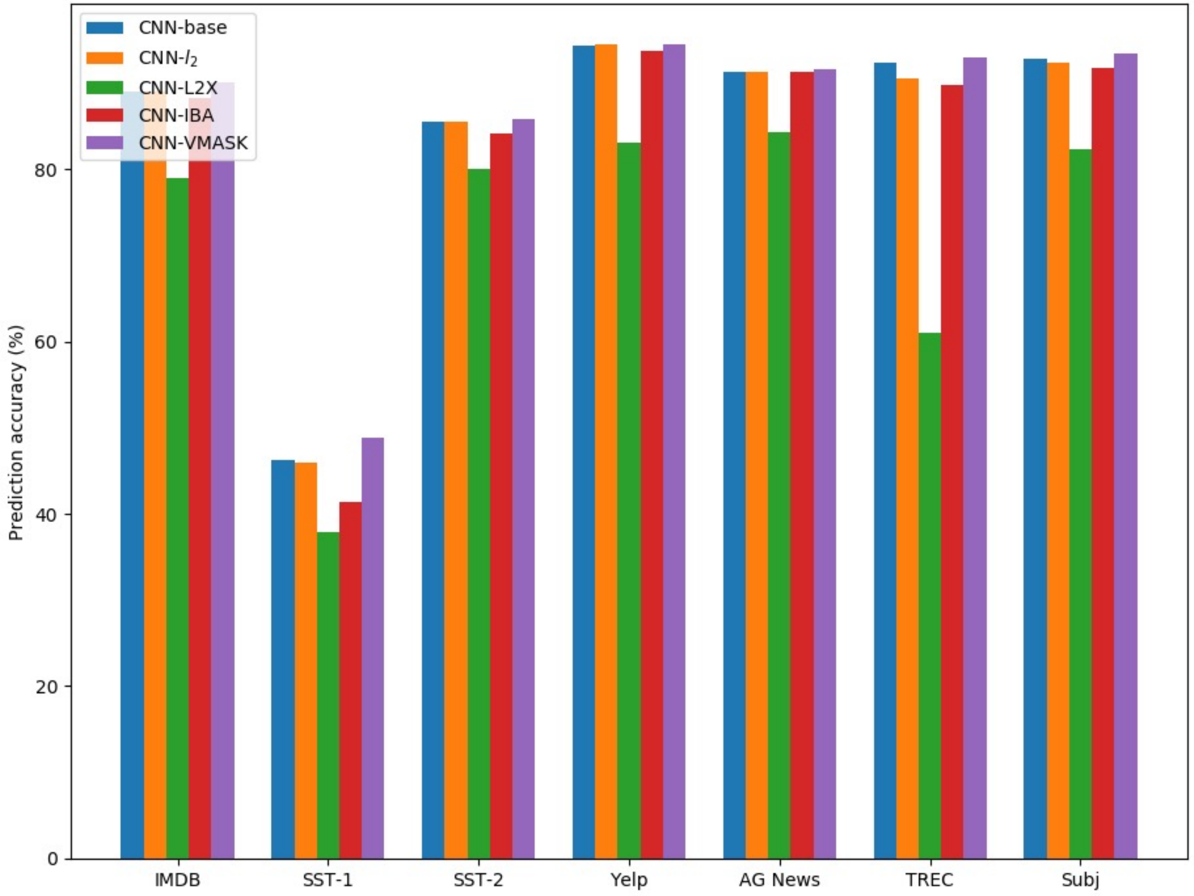
Top important words are sentiment words

Importance scores of irrelevant high-frequency words are around 0.5

Experiments

➢ Prediction accuracy (%)

VMASK improves model prediction accuracy



VMASK can help improve model generalization power

Experiments

➢ Local interpretability: AOPC

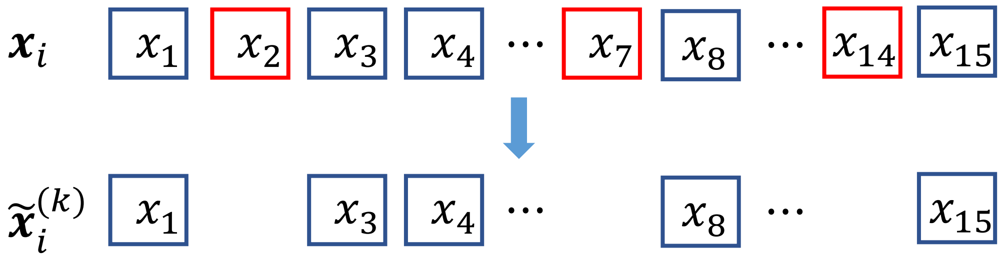
If a model is more interpretable, the post-hoc explanations would be more faithful

- Compare base and VMASK models with two model-agnostic explanation methods—LIME and SampleShapley

- The area over perturbation curve (AOPC) metric evaluates the faithfulness of explanations

$$AOPC(k) = \frac{1}{N}\sum_{i=1}^{N}\left\{p(\hat{y}|\boldsymbol{x}_i) - p\left(\hat{y}\,\middle|\,\tilde{\boldsymbol{x}}_i^{(k)}\right)\right\}$$

✓ Higher AOPCs are better

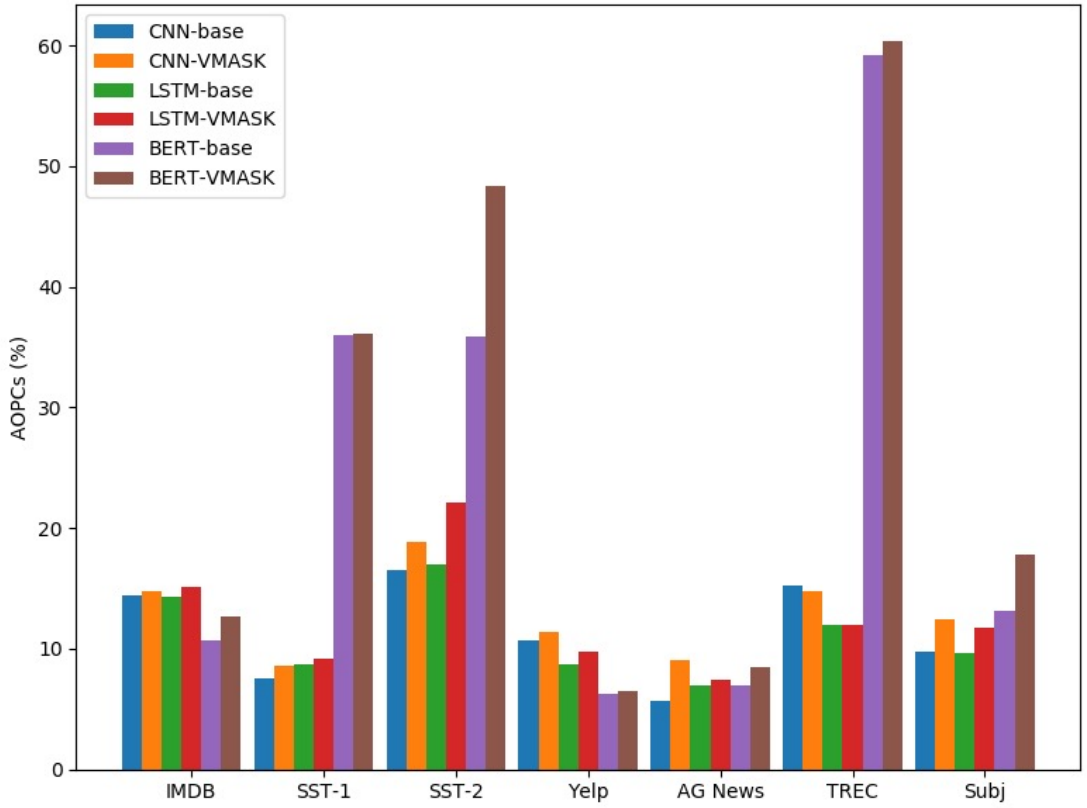


Experiments

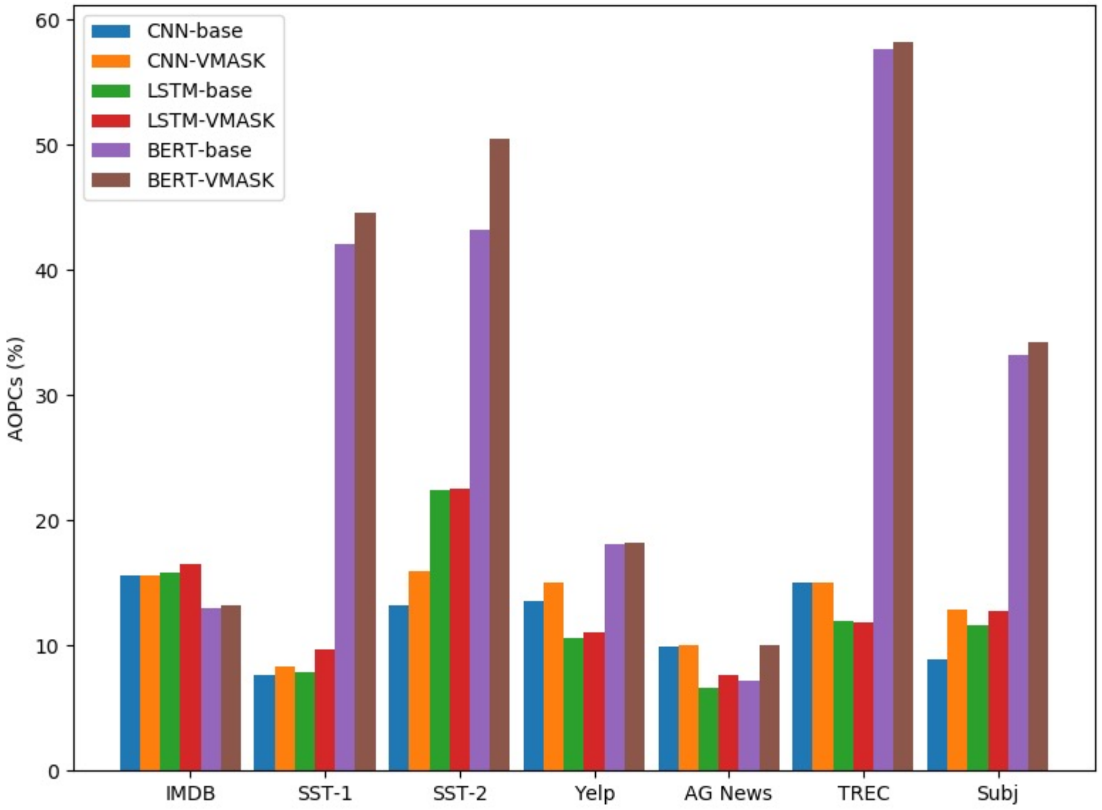
➢ AOPCs (%)

- The AOPCs of VMASK-based models are better
- VMASK can improve model's interpretability to post-hoc explanations

LIME



SampleShapley



Experiments

➢ Global interpretability: Post-hoc accuracy

- Global importance of words

$$\{\mathbb{E}[q(R_{x_t}|\boldsymbol{x}_t)]\}$$

- Evaluate the influence of global important words on model predictions

$$\text{Post-hoc-acc}(k) = \frac{1}{M}\sum_{m=1}^{M} 1[y_m(k) = y_m]$$

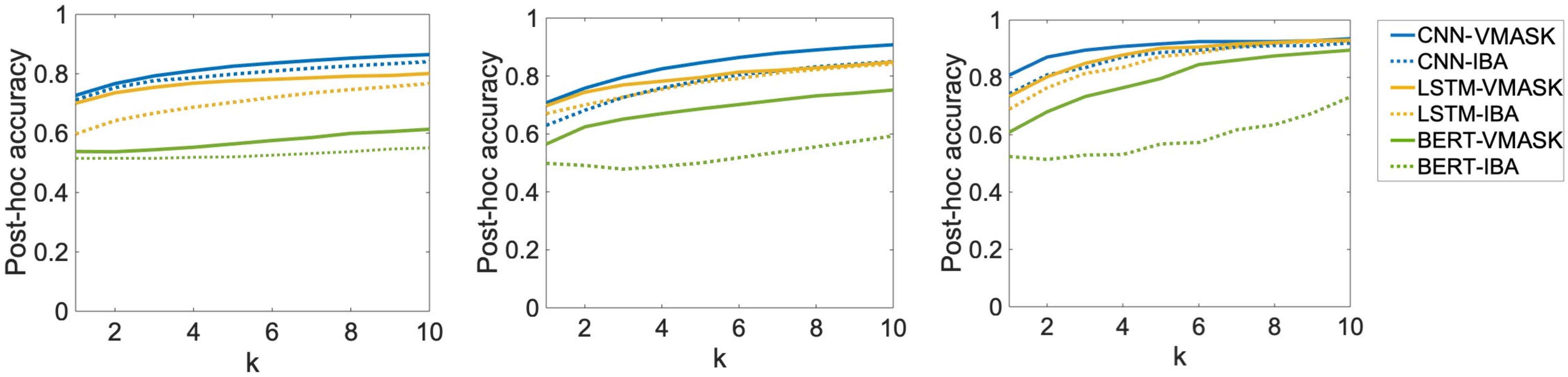
Top k words based on
global importance scores

✓ The higher, the better

Experiments

➢ Post-hoc accuracy (VMASK vs. IBA)

- VMASK is better on capturing task-specific important features than IBA
- BERT tends to use larger context with its self-attentions for predictions



IMDB　　Yelp　　Subj

Experiments

- Visualizing post-hoc local explanations
  - LIME explanations for different models on the IMDB dataset
  - For VMASK models, LIME can capture the sentiment words corresponding to the prediction

Models	Texts & Explanations	Prediction
CNN-base	Primary plot , primary direction , poor interpretation .	negative
CNN-VMASK	Primary plot , primary direction , poor interpretation .	negative
LSTM-base	John Leguizamo 's freak is one of the funniest one man shows I 've ever seen . I recommend it to anyone with a good sense of humor .	positive
LSTM-VMASK	John Leguizamo 's freak is one of the funniest one man shows I 've ever seen . I recommend it to anyone with a good sense of humor .	positive
BERT-base	Great story , great music . A heartwarming love story that ' s beautiful to watch and delightful to listen to . Too bad there is no soundtrack CD .	positive
BERT-VMASK	Great story , great music . A heartwarming love story that ' s beautiful to watch and delightful to listen to . Too bad there is no soundtrack CD .	positive

Experiments

➢ Visualizing post-hoc global explanations

- Adopt SP-LIME (Ribeiro et al., 2016) as a third-party to evaluate global interpretability
- Compute post-hoc global importance by summing all local importance scores of a feature (obtained from LIME local explanations)
- Compare base and VMASK-based models on the IMDB dataset

Models	Words
CNN-base	excellent, performances, brilliant
CNN-VMASK	excellent, fine, favorite
LSTM-base	plot, excellent, liked
LSTM-VMASK	excellent, favorite, brilliant
BERT-base	live, butcher, thrilling
BERT-VMASK	power, thrilling, outstanding

Irrelevant words

Experiments

➢ Visualizing post-hoc global explanations

- Adopt SP-LIME (Ribeiro et al., 2016) as a third-party to evaluate global interpretability
- Compute post-hoc global importance by summing all local importance scores of a feature (obtained from LIME local explanations)
- Compare base and VMASK-based models on the IMDB dataset

Models	Words
CNN-base	excellent, performances, brilliant
CNN-VMASK	excellent, fine, favorite
LSTM-base	plot, excellent, liked
LSTM-VMASK	excellent, favorite, brilliant
BERT-base	live, butcher, thrilling
BERT-VMASK	power, thrilling, outstanding

Sentiment words

Question?

Reference

- Camburu, Oana-Maria, et al. "e-snli: Natural language inference with natural language explanations." *Advances in Neural Information Processing Systems* 31 (2018).
- Chen, Hanjie, and Yangfeng Ji. "Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.
- Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. CoRR, abs/1508.05326.
- McCoy, R. Thomas, Ellie Pavlick, and Tal Linzen. "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference." *arXiv preprint arXiv:1902.01007* (2019).
- Poliak, Adam, et al. "Hypothesis only baselines in natural language inference." *arXiv preprint arXiv:1805.01042* (2018).
- Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. arXiv preprint physics/0004057.