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

Hanjie Chen, Yangfeng Ji

19 October 2020

Presenter: Sanchit Sinha

https://qdata.github.io/deep2Read/

Motivation

- Interpretability is important for reaffirming - reliability and trustworthiness of models
- Overcoming “black-box” nature of deep learning models
- Designing a model agnostic explainability method which can be used as a small addition during training
- It should be generalisable enough for all kinds of models, so only applied at the input layers
- No need for human intervention in providing correct or “ground truth” explanations or annotations
- Making interpretability one of the fundamental property of the network - “built into it”

Background

- Information Bottleneck framework

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

- Mutual information/Information Gain:

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_{(X,Y)}(x,y) \log\left(\frac{p_{(X,Y)}(x,y)}{p_X(x)\, p_Y(y)}\right),$$

Related Work

- Information Theory based explainability methods:
  - Maximizing mutual information to recognize important features - Chen et al., 2018; Guan et al., 2019
  - Optimizing the information bottleneck to identify feature attributions - Schulz et al., 2020; Bang et al., 2019
- Improving prediction using interpretability:
  - Post-hoc explanations to regularize models on prediction behaviors and force them to emphasize more on predefined important features - Ross et al., 2017; Ross and Doshi-Velez, 2018; Liu and Avci, 2019; Rieger et al., 2019
- Trying to correspond interpretability with human explainations
  - Camburu et al., 2018; Du et al., 2019b; Chen and Ji, 2019; Erion et al., 2019; Plumb et al., 2019

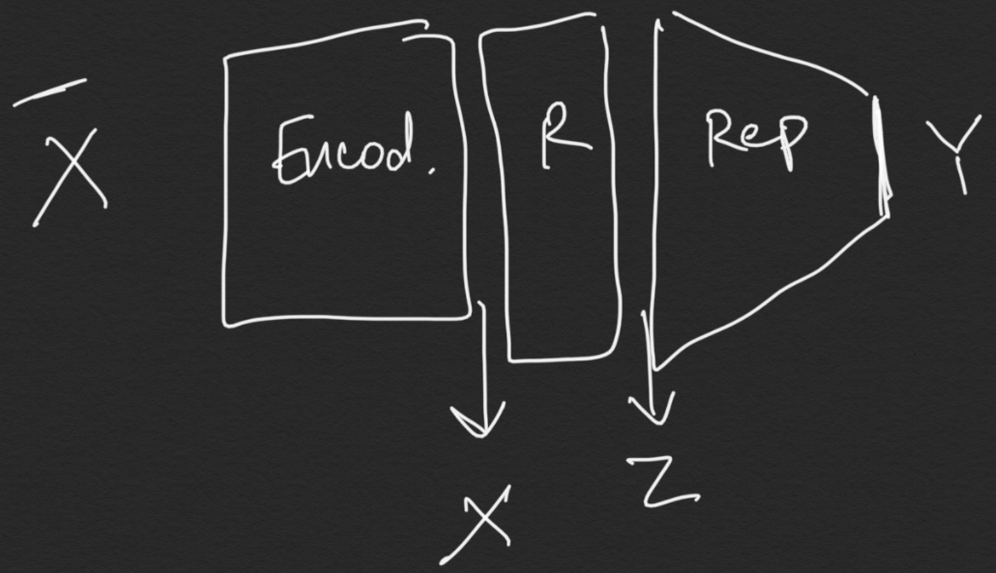
Claim / Target Task

- Introduce a model agnostic binary layer (mask) on top of the inputs which improves interpretability
- The layer acts as a filter to control the interactions of non-important words and only allows the important words to affect the output
- Layer gives better prediction and interpretability performance

Data Summary

Datasets	*C*	*L*	*#train*	*#dev*	*#test*
IMDB	2	268	20K	5K	25K
SST-1	5	18	8544	1101	2210
SST-2	2	19	6920	872	1821
Yelp	2	138	500K	60K	38K
AG News	4	32	114K	6K	7.6K
TREC	6	10	5000	452	500
Subj	2	23	8000	1000	1000

An Intuitive Figure Showing WHY Claim



$$X = E(\bar{X})$$

$$Z = R(X)$$

$$Y = Rep(Z)$$

$$\max \; I(Z,Y) - \beta I(Z,X)$$

Proposed Solution

- Notation :
  - X: Input words encoded into embeddings
  - R: Our custom V-Mask Layer. Every element is {0,1}
  - Z: Output of hadamard product of X and R
  - Y: The output prediction of the network (after classification)

- Main optimization function:

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

- Replacing p(x,y,z) [True distribution] with q(x,y,z) [Approximation distribution]

Proposed Solution

- Term-1:

$$
\begin{aligned}
I(\boldsymbol{Z};\boldsymbol{Y}) &\geq \sum_{\boldsymbol{y},\boldsymbol{z}} q(\boldsymbol{y},\boldsymbol{z}) \log p(\boldsymbol{y}|\boldsymbol{z}) + H_q(\boldsymbol{Y}) \\
&= \sum_{\boldsymbol{y},\boldsymbol{z},\boldsymbol{x}} q(\boldsymbol{x},\boldsymbol{y}) q(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{y}|\boldsymbol{z}) \\
&\quad + H_q(\boldsymbol{Y}), \qquad (3)
\end{aligned}
$$

$$
\begin{aligned}
I(\boldsymbol{Z};\boldsymbol{Y}^{(i)}) &\geq \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\boldsymbol{x}^{(i)}) \log p(\boldsymbol{y}^{(i)}|\boldsymbol{z}) \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}^{(i)})}[\log p(\boldsymbol{y}^{(i)}|\boldsymbol{z})].
\end{aligned}
$$

- Term-2:

$$
\begin{aligned}
I(\boldsymbol{Z};\boldsymbol{X}) &\leq \mathbb{E}_{q(\boldsymbol{x})}[\mathrm{KL}[q(\boldsymbol{z}|\boldsymbol{x}) \| p_0(\boldsymbol{z})]] \\
&= \mathrm{KL}[q(\boldsymbol{z}|\boldsymbol{x}^{(i)}) \| p_0(\boldsymbol{z})],
\end{aligned}
$$

- Final optimization:

$$
\begin{aligned}
\mathcal{L} = &\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}^{(i)})}[\log p(\boldsymbol{y}^{(i)}|\boldsymbol{z})] \\
&- \beta \cdot \mathrm{KL}[q(\boldsymbol{z}|\boldsymbol{x}^{(i)}) \| p_0(\boldsymbol{z})].
\end{aligned}
$$

Proposed Solution

- Substituting for R, we get

$$\begin{aligned}\mathcal{L} =&\mathbb{E}_{q(\boldsymbol{r}|\boldsymbol{x}^{(i)})}[\log p(\boldsymbol{y}^{(i)}|\boldsymbol{R}, \boldsymbol{x}^{(i)})] \\ &- \beta \cdot \mathrm{KL}[q(\boldsymbol{R}|\boldsymbol{x}^{(i)})||p_0(\boldsymbol{R})].\end{aligned}$$

- Tricks for optimization
  - Mean-field approximation (independence of rand vars)
  - Gumbel-softmax trick (for softmax)
  - KL cost annealing (posterior collapse)
- Models used - same as last paper CNN, LSTM, BERT

Experimental Results - Training

Models	Methods	IMDB	SST-1	SST-2	Yelp	AG News	TREC	Subj
CNN	CNN-base	89.06	46.32	85.50	94.32	91.30	92.40	92.80
	CNN-ℓ_2	89.12	46.01	85.56	94.46	91.28	90.62	92.39
	CNN-L2X	78.94	37.92	80.01	83.14	84.36	61.00	82.40
	CNN-IBA	88.31	41.40	84.24	93.82	91.37	89.80	91.80
	CNN-VMASK	**90.10**	**48.92**	**85.78**	**94.53**	**91.60**	**93.02**	**93.50**
LSTM	LSTM-base	88.39	43.84	83.74	95.06	91.03	90.40	90.20
	LSTM-ℓ_2	88.40	43.91	83.36	95.00	91.09	90.20	89.10
	LSTM-L2X	67.45	36.92	75.45	77.12	77.53	46.00	81.80
	LSTM-IBA	88.48	42.99	83.53	94.74	91.14	85.40	89.50
	LSTM-VMASK	**90.07**	**44.12**	**84.35**	**95.41**	**92.19**	**90.80**	**91.20**
BERT	BERT-base	91.80	53.43	92.25	96.42	93.59	96.40	95.10
	BERT-ℓ_2	91.75	52.08	92.25	96.41	93.52	96.80	94.80
	BERT-L2X	71.75	39.23	74.03	87.14	82.59	93.20	86.10
	BERT-IBA	91.66	53.80	92.24	96.27	93.45	96.80	95.60
	BERT-VMASK	**93.04**	**54.53**	**92.26**	**96.80**	**94.24**	**97.00**	**96.40**

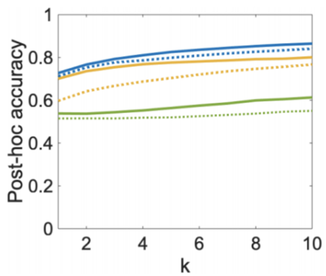
Experimental Results - AOPC

Methods	Models	IMDB	SST-1	SST-2	Yelp	AG News	TREC	Subj
LIME	CNN-base	14.47	7.59	16.50	10.69	5.66	15.28	9.77
	CNN-VMASK	**14.74**	**8.63**	**18.86**	**11.38**	**9.03**	14.81	**12.40**
	LSTM-base	14.34	8.76	17.03	8.72	7.00	11.95	9.67
	LSTM-VMASK	**15.10**	**9.52**	**22.14**	**9.70**	**7.39**	11.97	**11.68**
	BERT-base	10.63	36.00	35.89	6.30	7.00	59.22	13.08
	BERT-VMASK	**12.64**	36.16	**46.87**	6.49	**8.47**	**60.37**	**17.82**
SampleShapley	CNN-base	15.53	7.63	13.15	13.57	9.88	14.97	8.84
	CNN-VMASK	15.53	**8.33**	**15.95**	**15.06**	9.98	**15.03**	**12.88**
	LSTM-base	15.80	7.91	22.38	10.55	6.62	11.90	11.66
	LSTM-VMASK	**16.48**	**9.73**	22.52	**10.99**	**7.65**	11.86	**12.74**
	BERT-base	12.97	42.06	43.16	18.06	7.21	57.69	33.22
	BERT-VMASK	**13.18**	**44.57**	**50.44**	18.17	**10.02**	**58.26**	**34.22**

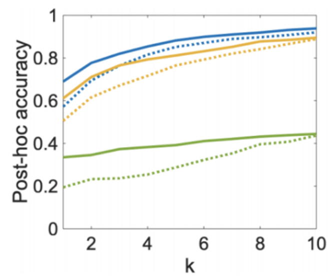
Experimental Results - Post-hoc accuracy

- Influence of top k% words on the accuracy vs the whole text

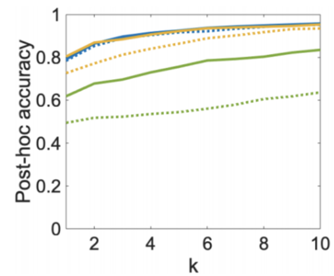
$$\text{post-hoc-acc}(k) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}[y_m(k) = y_m],$$



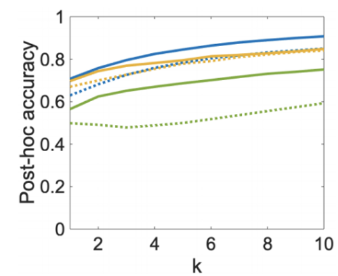
(a) IMDB



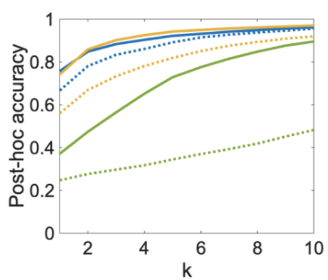
(b) SST-1



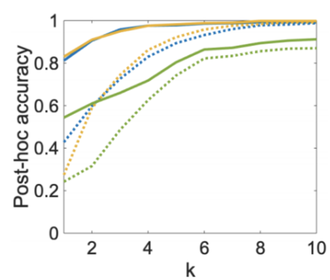
(c) SST-2



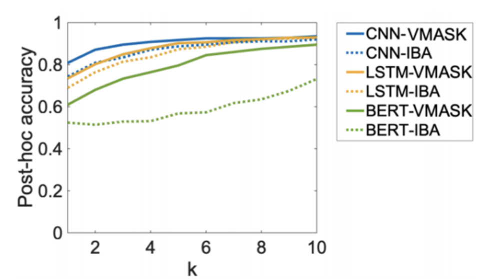
(d) Yelp



(e) AG News



(f) TREC



(g) Subj

Experimental Results - Qualitative

Models	Texts	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