Attending via both Fine-tuning and Compressing

Jie Zhou[1,2], Yuanbin Wu[2], Qin Chen[2], Xuanjing Huang[3], and Liang He[1,2]

[1]Shanghai Key Laboratory of Multidimensional Information Processing
[2]School of Computer Science and Technology, East China Normal University
jzhou@ica.stc.sh.cn, {ybwu,qchen,lhe}@cs.ecnu.edu.cn
[3]School of Computer Science, Fudan University
xjhuang@fudan.edu.cn

Abstract

Though being a primary trend for enhancing interpretability of neural networks, attention mechanism's reliability and validity are still under debate. In this paper, we try to purify attention scores to obtain a more faithful explanation of downstream models. Specifically, we propose a framework consisting of a learner and a compressor, which performs fine-tuning and compressing iteratively to enhance the performance and interpretability of the attention mechanism. The learner focuses on learning better text representations to achieve good decisions by fine-tuning, while the compressor aims to perform compressions over the representations to retain the most useful clues for explanations with a Variational information bottleneck ATtention (VAT) mechanism. Extensive experiments on eight benchmark datasets show the great advantages of our proposed approach in terms of both performance and interpretability.

1 Introduction

Attention mechanisms (Bahdanau et al., 2014) have achieved great success in various natural language processing (NLP) tasks. They are introduced to mimic the human eye focusing on important parts in the inputs when predicting labels. The existing studies show attention mechanisms can improve not only the performance but also the interpretability of the models (Mullenbach et al., 2018; Xie et al., 2017; Xu et al., 2015). Li et al. (2016) pointed the view: "Attention provides an important way to explain the workings of neural models". Additionally, Wiegreffe and Pinter (2019) showed that attention mechanisms could help understand the inner workings of a model.

The basic assumption of understanding of models with attention scores is that the inputs (e.g., words) with high attentive weights are essential for making decisions. However, as far as we know, it has not been formally verified. Existing research (Jain and Wallace, 2019) also shows that attention is not explicable, and there are a lot of controversy regarding to the result explanations (Wiegreffe and Pinter, 2019; Jain and Wallace, 2019). Moreover, we find that though the attention mechanism can help improve the performance for text classification in our experiments, it may focus on the irrelevant information. For example, in the sentence "A very funny movie.", the long short-term memory model with standard attention (LSTM-ATT) infers a correct sentiment label while pays more attention to the irrelevant word "movie", making the result difficult to explain.

In general, the attention weights are only optimized to encode the task-relevant information while are not restricted to imitate human behavior. In order to enhance the interpretability of the attention mechanism, recent studies turn to integrate the human provided explanation signals into the attention models. Rei and Søgaard (2018) regularized the attention weights with a small amount of word-level annotations. Barrett et al. (2018); Bao et al. (2018) improved the explanation of attention by aligning explanations with human-provided rationales. These methods rely on additional labour consuming labelling for enhancing explanations, which is hard to extend to other datasets or tasks.

In this paper, we aim to train a more efficient and effective interpretable attention model without any pre-defined annotations or pre-collected explanations. Specifically, we propose a framework consisting of a learner and a compressor, which enhances the performance and interpretability of the attention model for text classification[1]. The learner learns text representations by fine-tuning

[1]We focus on the task of text classification, but our method can be easily extended to other NLP or CV tasks with attention mechanisms.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2152–2161

the encoder. Regarding to the compressor, we are motivated by the effectiveness of the information bottleneck (IB) (Tishby et al., 1999) to enhance performance (Li and Eisner, 2019) or detect important features (Bang et al., 2019; Chen and Ji, 2020; Jiang et al., 2020; Schulz et al., 2020), and present a Variational information bottleneck ATtention (VAT) mechanism using IB to keep the most relevant clues and forget the irrelevant ones for better attention explanations. In particular, IB is integrated into attention to minimize the mutual information (MI) with the input while preserving as much MI as possible with the output, which provides more accurate and reliable explanations by controlling the information flow.

To evaluate the effectiveness of our proposed approach, we adapt two advanced neural models (LSTM and BERT) within the framework and conduct experiments on eight benchmark datasets. The experimental results show that our adapted models outperform the standard attention-based models over all the datasets. Moreover, they exhibit great advantages with respect to interpretability by both qualitative and quantitative analyses. Specifically, we obtain significant improvements by applying our model to the semi-supervised word-level sentiment detection task, which detects the sentiment words based on attention weights via only sentence-level sentiment label. In addition, we provide the case studies and text representation visualization to have an insight into how our model works.

The main contributions of this work are summarized as follows.

- We propose a novel framework to enhance the performance and interpretability of the attention models, where a learner is used to learn good representations by fine-tuning and a compressor is used to obtain good attentive weights by compressing iteratively.
- We present a Variational information bottleneck ATtention (VAT) mechanism for the compressor, which performs compression over the text representation to keep the task related information while reduce the irrelevant noise via information bottleneck.
- Extensive experiments show the great advantages of our models within the proposed framework, and we perform various qualitative and quantitative analyses to shed light on why our models work in both performance and interpretability.

2 Related Work

In this section, we survey related attention mechanisms (Bahdanau et al., 2014) and review the most relevant studies on information bottleneck (IB) (Tishby et al., 1999).

Attention has been proved can help explain the internals of neural models (Li et al., 2016; Wiegreffe and Pinter, 2019) though it is limited (Jain and Wallace, 2019). Many researchers try to improve the interpretability of the attention mechanisms. Rei and Søgaard (2018) leveraged small amounts of word-level annotations to regularize attention. Kim et al. (2017) introduced a structured attention mechanism to learn attention variants from explicit probabilistic semantics. Barrett et al. (2018); Bao et al. (2018) aligned explanations with human-provided rationales to improve the explanation of attention. Unlike these methods that require prior attributions or human explanations, the VAT method enforces the attention to learn the vital information while filter the noise via IB.

A series of studies motivate us to utilize IB to improve the explanations of attention mechanisms. Li and Eisner (2019) compressed the pre-trained embedding (e.g., BERT, ELMO), remaining only the information that helps a discriminative parser through variational IB. Zhmoginov et al. (2019) utilized the IB approach to discover the salient region. Some works (Jiang et al., 2020; Chen et al., 2018; Guan et al., 2019; Schulz et al., 2020; Bang et al., 2019) proposed to identify vital features or attributions via IB. Moreover, Chen and Ji (2020) designed a variational mask strategy to delete the useless words in the text. As far as we are aware, we are the first ones to leverage IB into attention mechanisms to train more interpretable attention with better accuracy.

3 Our Approach

In this section, we introduce our framework consisting of a learner and a compressor with a Variational information bottleneck ATtenttion (VAT) mechanism. Given an attention-based neural network model, we formulate our idea within the framework of variational information bottleneck (VIB) (Tishby et al., 1999). Our framework aims to improve the attention's interpretalility with better performance by restricting the attention to capture the crucial words while filter the useless information.

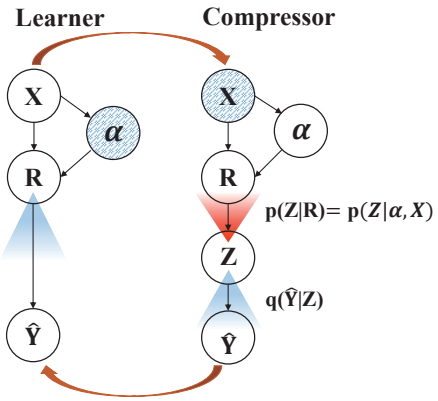


Figure 1: The framework.The learner aims to learn the good text representation X by fine-tuning, and the compressor aims to learn good attention weights by compressing the attentive representations to capture the important words while forget the redundant information via VAT. The blue circles mean the corresponding parameters of the modules are fixed.

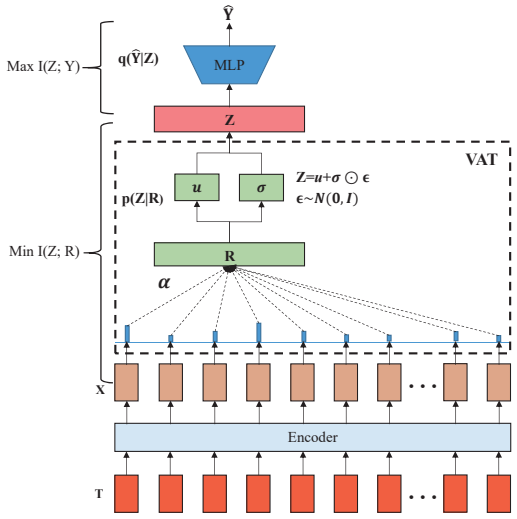


Figure 2: The architecture of our VAT (Compressor). First, we obtain the input text's word representations X via an encoder trained by the learner. Then, we calculate Z by compressing the text representation R that is the weighted sum of X based on the attention α, while remaining the maximum information to judge Y by inputting Z into a MLP classifier for predicting.

3.1 Overview

Our framework is composed of a learner and a compressor, which performs fine-tuning and compressing iteratively (Figure 1). The learner aims to learn a task-specific contextual word representation by fine-tuning. The compressor enforces the model to learn task-relevant information while reduce irrelevant information via IB. We iteratively perform the learner and compressor (fine-tuning and compressing) to improve each other.

Learner. We adopt a basic attention-based neural network model as a learner to learn representations of the words based on the good attention weights learned by the compressor. The model is optimized by cross-entropy loss to learn the label-relevant information. In this phase, we fix the attention's parameters so that the model will focus on updating the encoder to learn word representations.

Compressor. To restrict the attention to capture the vital information while reduce the noise, we integrate IB into attention mechanisms to compress the text attentive representation. We fix the encoder's parameters so that the model will focus on learning the attention weights based on current representations obtained from the learner.

3.2 Basic Attention Model (Learner)

In this section, we describe our learner, which is an attention-based neural network model. First, given a text $T = \{w_1, w_2, ..., w_{|T|}\}$, where $|T|$ is the length of text T, we feed it into an encoder with a word embedding layer. We adopt LSTM and BERT models as our encoder, and other models can also be applied to our framework. We obtain the context-aware word representations $x = [x_1, x_2, ..., x_{|T|}]$, where x_i is the hidden vector of the word w_i.

$$x = encoder(T, \theta_{encoder}), \tag{1}$$

where $\theta_{encoder}$ is the parameters of the encoder.

Based on the contextual word representations, attention mechanism (Bahdanau et al., 2014) [2] is utilized to capture the important parts in the text and obtain the text representation R, which is calculated as,

$$\begin{aligned} R &= \sum_{i=1}^{n} \alpha_i x_i \\ \alpha_i &= \text{softmax}(\mathbf{v}_a^\top \tanh(\mathbf{W}_a x_i)) \end{aligned} \tag{2}$$

where $\theta_{attention} = \{v_a, W_a\}$ is the trainable parameters of the attention, which is not updated in this step to learn the word representation x based the good attention learned by the compressor. $\alpha = [\alpha_1, \alpha_2, ..., \alpha_{|T|}]$ is the attention weights. Finally, we input the text representation R into a

[2] In this paper, we only explore the local attention mechanism on our framework, other attention mechanisms (e.g., multi-head attention (Vaswani et al., 2017)) can also be applied. We would like to explore it in future work.

multi-layer perceptron (MLP) to predict the probability. The cross-entropy loss is used to optimize the model.

3.3 Variational Information Bottleneck Attention (Compressor)

The learner optimizes the sentence representations by minimizing the cross-entropy loss, which does not restrict the model to ignore the useless information. Thus, we compress sentence representations R into a latent representation Z that retains most useful information to infer the label Y. We propose to accomplish this by integrating VIB into the attention mechanism (Figure 2).

To ensure Z contains maximum ability to predict Y ($I(Z;Y)$) while has the least redundant information form R ($-I(Z;R)$), we use the standard IB theory (Tishby et al., 1999) and define the objective function as:

$$\max_{\boldsymbol{\alpha}} I(Z;Y) - \beta \cdot I(Z;R) \tag{3}$$

where $I(\cdot;\cdot)$ means the mutual information and β is a coefficient to balance two components. The main challenge is to estimate the lower bound for $I(Z;Y)$ and the upper bound for $I(Z;R)$. [3]

The joint probability $p_\theta(r,y,z)$ can be factored as $p(r)\cdot p(y \mid r)\cdot p_\theta(z \mid r)$ based on the independence assumption [4]. By replacing the conditional distribution $p_\theta(y \mid z)$ with a variational approximation $q_\phi(y \mid z)$, we obtain a lower bound of $I(Z;Y)$. $q_\phi(y \mid z)$ is a simple classifier that runs on a compressed text representation z.

$$\begin{aligned}&\overbrace{\mathbb{E}_{p_\theta(y,z)}[\log \frac{p_\theta(y \mid z)}{p(y)}]}^{I(Z;Y)} - \overbrace{\mathbb{E}_{p_\theta(y,z)}[\log \frac{q_\phi(y \mid z)}{p(y)}]}^{\text{lower bound}} \\ &= \mathbb{E}_{p_\theta(z)}[\mathrm{KL}(p_\theta(y \mid z) \| q_\phi(y \mid z))] \geqslant 0\end{aligned} \tag{4}$$

where $KL[\cdot\|\cdot]$ represents Kullback-Leibler divergence.

Specifically, we regard $p(y)$ as constant and then minimize $\mathbb{E}_{p_\theta(y,z)}[\log q_\phi(y \mid z)]$. Since we must first sample r to sample y, z from $p_\theta(r,y,z)$, the lower bound of $I(Z;Y)$ is computed as,

$$I(Z;Y) \geqslant \mathbb{E}_{p(r,y)}[\mathbb{E}_{p_\theta(z|r)}[\log q_\phi(y \mid z)]] \tag{5}$$

We calculate the upper bound of $I(Z;R)$ by replacing $p_\theta(z)$ with a variational distribution $r_\psi(z)$,

$$\begin{aligned}&\overbrace{\mathbb{E}_{p(r)}[\mathbb{E}_{p_\theta(z|r)}[\log \frac{p_\theta(z \mid r)}{r_\psi(z)}]]}^{\text{upper bound}} - \overbrace{\mathbb{E}_{p(r)}[\mathbb{E}_{p_\theta(z|r)}[\log \frac{p_\theta(z \mid r)}{p(z)}]]}^{\mathrm{I}(Z;R)} \\ &= \mathbb{E}_{p(r)}[\mathrm{KL}(p(z) \| r_\psi(z))] \geqslant 0\end{aligned} \tag{6}$$

The upper bound of $I(Z;R)$ is computed as,

$$\begin{aligned}I(Z;R) &\leqslant \mathbb{E}_{p(r)}[\mathbb{E}_{p_\theta(z|r)}[\log \frac{p_\theta(z \mid r)}{r_\psi(z)}]] \\ &= \mathbb{E}_{p(r)}[\mathrm{KL}[p_\theta(z \mid r) \| r_\psi(z)]]\end{aligned} \tag{7}$$

Then, we obtain the lower bound $\mathcal{L}$ of IB by substituting Equation 5 and 7 into Equation 3:

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{p(r,y)}[&\mathbb{E}_{p_\theta(z|r)}[\log q_\phi(y \mid z)] \\ &- \beta \cdot \mathrm{KL}[p_\theta(z \mid r) \| r_\psi(z)]]\end{aligned} \tag{8}$$

The first component in $\mathcal{L}$ is to keep the most useful information in $p_\theta(z|r)$ for inferring y, while the second one is to regularize $p_\theta(z|r)$ with a predefined prior distribution $r_\psi(z)$ (e.g., Gaussian distribution). To compute $p_\theta(z|r)$, we adopt the reparametrization trick for multivariate Gaussians (Rezende et al., 2014), which obtains the gradient of parameters that derive z from a random noise ϵ.

$$z = u + \sigma \odot \epsilon, \epsilon \sim N(0, I) \tag{9}$$

where $\odot$ means element-wise multiplication. u and σ denote the mean and covariance defined by two functions of R, where $R = \alpha \cdot x$ that is learned based on attention. In particular, two MLP are used to predict u and σ.

Finally, we input the z into a MLP to predict $q_\phi(y \mid z)$ and optimize the attention's parameter via Equation 8.

4 Experiment Setup

We adopt two typical neural network models, attention-based LSTM (Hochreiter and Schmidhuber, 1997) and BERT (Devlin et al., 2019), to explore our VAT algorithm.

4.1 Datasets and Baselines

Datasets To evaluate the effectiveness of our VAT model, we conduct the experiments over eight benchmark datasets: IMDB (Maas et al., 2011), Stanford Sentiment Treebank with (includes SST-1 and its binary version SST-2) (Socher et al., 2013), Yelp (Zhang et al., 2015), AG News (Zhang et al., 2015), TREC (Li and Roth, 2002), subjective/objective classification Subj (Pang and Lee, 2005) and Twitter (Rosenthal et al., 2015, 2014). The statistics information of these datasets are shown in Table 1.

[3] We give the main steps as follows and the detailed derivation is provided in supplementary materials.

[4] $Y \rightarrow R \rightarrow Z$: Y and Z are independent given R.

	IMDB	SST-1	SST-2	Yelp	AG News	Trec	Subj	Twitter
Class	2	5	2	2	4	6	2	3
Length	268	18	19	138	32	10	23	22
#train	20,000	8,544	6,920	500,000	114,000	5,000	8,000	7,969
#dev	5,000	1,1101	872	60,000	6,000	452	1,000	1,375
#test	25,000	2,210	1,821	38,000	7,600	500	1,000	3,795

Table 1: The statistics information of the datasets, where Class is the number of the class, Length is average text length, and #train/#dev/#test counts the number of samples in the train/dev/test sets.

	IMDB	SST-1	SST-2	Yelp	AG News	Trec	Subj	Twitter	Average
LSTM-base	88.79	45.20	85.45	95.10	91.91	90.00	89.00	71.25	82.09
LSTM-ATT	88.16	46.29	84.73	95.06	91.88	91.00	90.80	70.75	82.33
LSTM-VAT	**88.98**	**47.42**	**86.22**	**95.32**	**92.04**	**92.80**	**91.10**	**71.62**	**83.19**
BERT-base	91.90	51.44	91.60	96.07	93.52	96.60	96.50	75.28	86.61
BERT-ATT	91.81	51.13	91.16	97.20	93.41	96.40	96.20	74.84	86.52
BERT-VAT	**92.11**	**51.99**	**91.98**	**97.36**	**93.71**	**97.20**	**96.70**	**77.13**	**87.27**

Table 2: The main results of text classification.

Baselines We compare our model with two kinds of models, basic models (LSTM/BERT-base) and attention-based models (LSTM/BERT-ATT). LSTM-base takes the max-pooling of the LSTM's hidden vectors as text representation. For BERT-base, the "[CLS]" representation is obtained as the sentence representation. LSTM-ATT model is a standard attention-based LSTM model that has the same structure as the learner. We obtain the BERT-ATT by replacing the LSTM encoder with BERT in LSTM-ATT. Our models are marked with VAT (LSTM-VAT, BERT-VAT), which integrate VIB into attention-based neural models.

4.2 Implementation Details

For LSTM-based models, we use GloVe embedding (Pennington et al., 2014) with 300-dimension to initialize the word embedding and fine-tune it during the training. We randomly initialize all out-of-vocabulary words and weights with the uniform distribution $U(-0.1, 0.1)$. For the BERT-based models, we fine-tune pre-trained BERT-base model. The dimension of hidden state vectors of LSTM is 100 and the max sentence length is 256 in our experiments. Adam (Kingma and Ba, 2014) is utilized as the optimizer with learning rate 0.001 (for LSTM-based model) and 0.00001 (for BERT-based model). We also search different values $\beta \in \{0.01, 0.1, 1, 10\}$.

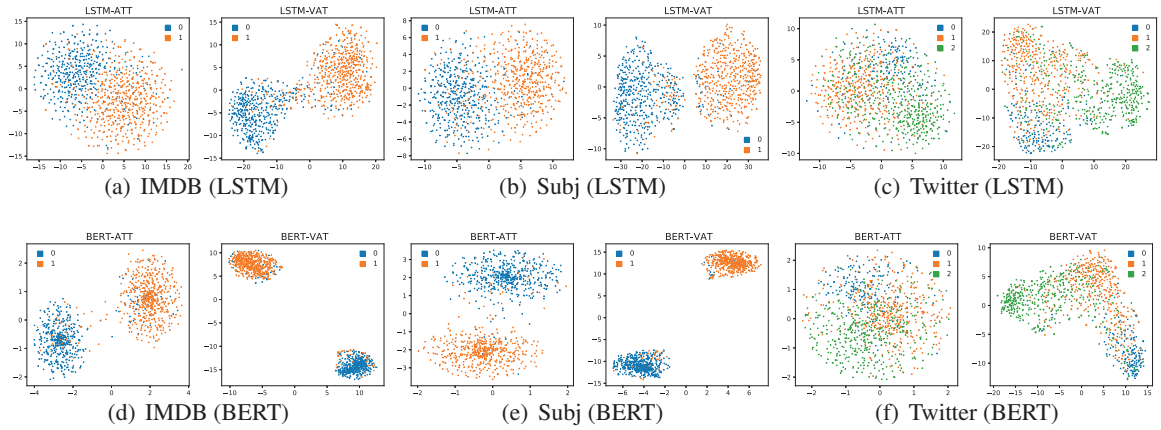
5 Experiments

First, we perform our models and baselines on eight benchmark datasets and visualize the text representation to verify the effectiveness of VAT (Section 5.1). Second, to further investigate our VAT model, we adopt two popular explanation metrics for quantitative evaluation (Section 5.2). Third, we apply our models to semi-supervision sentiment detection task to evaluate the explanation of our model (Section 5.3). Fourth, we explore the influence of our iteration strategy in Section 5.4 and provide case studies in Section 5.5. For the limitation of the space, we may only list the results on parts of the datasets in some cases since the conclusions are similar for other datasets. The complete results are presented in the supplementary materials.

5.1 Main Results

We report the accuracy of our VAT and baselines based on LSTM and BERT (Table 2). From these results, we find the following observations: **1)** our models (LSTM/BERT-VAT) outperform all the corresponding baselines over all the eight datasets, which denotes the effectiveness of our VAT on both LSTM and BERT-based models; **2)** compared with attention-based models (LSTM/BERT-ATT), our models obtain better results. It indicates reducing the irrelevant information in input via VAT can improve the performance of the models.

Furthermore, we visualize the sentence representations obtained from LSTM/BERT-ATT and -VAT models (Figure 3). We randomly select 1000 samples from the test set for each dataset. We can find that our VAT model can reduce the distance of the samples in a class and add the distance of the samples in different classes. For example, it is hard to split the positive samples from the negative ones based on the representations obtained from LSTM-ATT for the IMDB dataset, while the divider line based on our VAT is clear. These ob-



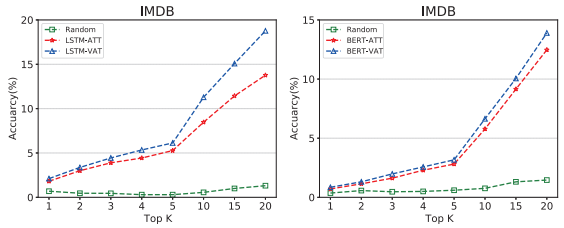
(a) IMDB (LSTM) (b) Subj (LSTM) (c) Twitter (LSTM)

(d) IMDB (BERT) (e) Subj (BERT) (f) Twitter (BERT)

Figure 3: Visualization of text representation obtained from LSTM/BERT-ATT and LSTM/BERT-VAT. We use t-SNE to transfer 100/768-dimensional feature space into two-dimensional space.

		IMDB	SST-1	SST-2	Yelp	AG News	Trec	Subj	Twitter
Accuracy	LSTM-base	88.79	45.20	85.45	95.10	91.91	90.00	89.00	71.25
AOPC	Random	0.30	5.97	7.58	1.02	1.87	19.40	1.50	4.72
	LSTM-ATT	5.27	12.94	20.54	6.64	5.99	31.00	2.10	19.10
	LSTM-VAT	**6.13**	**14.34**	**21.58**	**7.12**	**6.59**	**37.20**	**6.30**	**20.37**
Accuracy	BERT-base	91.90	51.44	91.60	96.07	93.52	96.60	96.50	75.28
AOPC	Random	0.60	33.26	41.46	3.60	44.20	65.80	45.70	59.21
	BERT-ATT	2.81	33.98	41.52	4.73	52.22	71.60	45.70	59.39
	BERT-VAT	**3.17**	**34.03**	**41.52**	**6.64**	**54.70**	**72.20**	**45.80**	**59.45**

Table 3: The results of AOPC.



(a) IMDB (LSTM) (b) IMDB (BERT)

Figure 4: The influence of Top-K for LSTM/BERT-based models in terms of AOPC.

servations show our VAT model can learn a better task-specific representation by enforcing the model to reduce the task-irrelevant information.

5.2 Quantitative Evaluation

In this section, we evaluate our VAT model using two metrics, AOPC and post-hoc accuracy, which are widely used for explanations (Chen and Ji, 2020). Note that well-trained LSTM/BERT-base is used for evaluating the performance of classification.

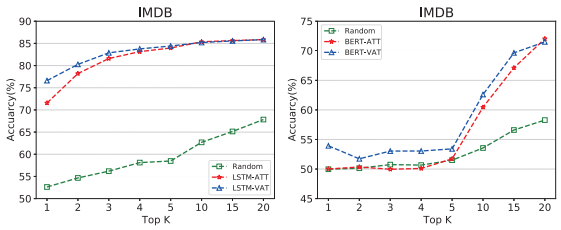
AOPC. To evaluate the faithfulness of explanations to our models, we adopt the area over the perturbation curve (AOPC) (Nguyen, 2018; Samek et al., 2016) metric. It calculates the average change of accuracy over test data by deleting top K words via attentive weights. The larger the value of AOPC, the better the explanations of the models.

Table 3 displays the results with $K = 5$. We compare our models with random and basic attention-based models. From the results, we observe that: **1)** basic attention-based models (LSTM/BERT-ATT) can find the important words in the sentence to some extent. Comparing with random (Random), LSTM/BERT-ATT obtains significant improvement; **2)** Our models (LSTM/BERT-VAT) outperform the standard attention-based models. It indicates that integrating VIB into the attention mechanism can help improve the interpretability of the models by filtering the useless information; **3)** BERT model is sensitive to the context; deleting the words will destroy the semantic information of the sentence and significantly affect the model's performance.

We also explore the influence of top-K (Figure 4). Intuitively, the more words we delete, the larger accuracy the models reduce. Our models reduce more performance than random and attention-based

		IMDB	SST-1	SST-2	Yelp	AG News	Trec	Subj	Twitter
Accuracy	LSTM-base	88.79	45.20	85.45	95.10	91.91	90.00	89.00	71.25
Post-hoc	Random	58.48	34.21	71.33	64.74	62.45	71.40	78.40	54.07
	LSTM-ATT	83.96	40.56	82.70	87.80	78.96	73.60	87.40	70.20
	LSTM-VAT	**84.41**	**43.39**	**84.35**	**88.82**	**81.43**	**79.20**	**89.10**	**71.23**
Accuracy	BERT-base	91.90	51.44	91.60	96.07	93.52	96.60	96.50	75.28
Post-hoc	Random	51.50	20.27	50.52	50.21	26.74	26.60	50.60	40.50
	BERT-ATT	51.72	29.19	58.92	53.63	37.53	34.20	61.90	53.68
	BERT-VAT	**53.40**	**30.23**	**61.34**	**56.58**	**43.08**	**36.80**	**65.40**	**56.05**

Table 4: The results of post-hoc accuracy.



(a) IMDB (LSTM) (b) IMDB (BERT)

Figure 5: The influence of Top-K for LSTM-based models in terms of post-hoc.

models. For the IMDB dataset, when deleting top 20 words (average length is 268), the accuracy reduces about 19 points for our LSTM-VAT model while it is about 2 points for the random model.

Post-hoc Accuracy. We also adopt the post-hoc accuracy (Chen et al., 2018) to evaluate the influence of task-specific essential words on the performance of LSTM-based and BERT-based models. For each test sample, we select the top K words based on their attentive weights as input to make a prediction and compare it with the ground truth.

Table 4 presents the performance with $K = 5$. **First**, it is interesting to find that the post-hoc accuracy with five most important words on Sbuj dataset (89.10) is even better than the original sentence (89.00). Additionally, we obtain comparable results with only five words for SST-1, SST-2, and Twitter datasets. These show that our model can reduce the noise information since most of the words are useless for predictions in some cases. **Second**, for BERT-based models, the context words are also important for classification even though they may not be task-specific.

Similarly, we investigate the influence of top-K for post-hoc (Figure 5). The LSTM-base model with top-10 words selected by our LSTM-VAT model can achieve comparable results with the original samples in most cases. Additionally, for the IMDB dataset, the accuracy of LSTM-base with one word selected by our VAT model is even better than the one with 20 words selected randomly.

5.3 Semi-Supervised Word-Level Sentiment Detection

We perform semi-supervised word-level sentiment detection in Twitter (Rosenthal et al., 2015, 2014) to evaluate the interpretability of our VAT. This task requires to detect the sentiment words in a tweet via the sentiment polarity of the whole tweet. In the following example from the dataset, positive words ("good" and "fantastic") are marked with a bold font and the overall polarity of the tweet is positive:

***Good** morning becky! Thursday is going to be **fantastic**!*

We use the SemEval 2013 Twitter dataset, which contains word-level sentiment annotation. We remove the samples with the neutral sentiment. We report word-level precision, recall, and F-measure for evaluating the models (Table 5), the same as (Rei and Søgaard, 2018). Note that we select the top-K (we set it as 1 and 5 here) words according to the attention weights as the sentiment words.

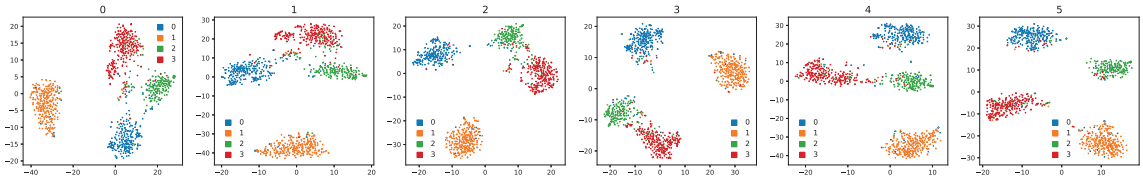
We compare our VAT model with random and attention-based models. The results show attention-based models can capture the important words in the text, to a certain extent. Since our VAT can reduce irrelevant information, it performs better than the standard attention model. Also, LSTM-based models outperform BERT-based models for this task in most cases. It is because that BERT learns much semantic information from the text, and context information plays a vital role in prediction.

5.4 Influence of Iteration

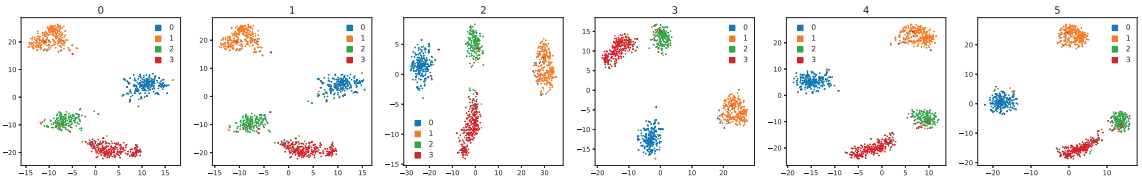
We propose to train the learner and compressor iteratively so that the learner optimizes the word representations based on the good attention, and the compressor optimizes the attention based on the good word representations. To have a deep look at how it works, we first provide our VAT model's accuracy with different iterations (Table 6). From the results, we can find that the model's performance will improve at first, then it will converge.

	Positive						Negative					
	P@1	R@1	F1@1	P@5	R@5	F1@5	P@1	R@1	F1@1	P@5	R@5	F1@5
Random	14.88	4.78	6.56	14.59	23.34	16.06	20.52	5.61	8.19	17.18	23.68	17.97
LSTM-ATT	58.70	26.04	32.73	30.30	54.17	34.70	47.13	15.74	21.39	28.24	42.04	30.33
LSTM-VAT	**65.20**	**29.38**	**36.60**	33.04	**58.40**	37.77	**60.00**	**21.42**	**28.76**	32.70	**49.19**	**35.35**
BERT-ATT	46.44	16.52	21.82	33.13	52.52	35.66	37.74	9.19	13.46	30.82	39.65	30.23
BERT-VAT	55.24	20.62	26.90	**37.26**	58.39	**40.09**	43.83	11.15	16.20	**36.42**	44.55	35.30

Table 5: The results of semi-supervision word-level sentiment detection in twitter.



(a) AG News (LSTM)



(b) AG News (BERT)

Figure 6: Visualization of text representation obtained from LSTM/BERT-VAT with different iterations. We use t-SNE to transfer 100/768-dimensional feature space into two-dimensional space.

Method	Text	Prediction
LSTM-ATT	I admired this work a lot.	Positive √
LSTM-VAT	I admired this work a lot.	Positive √
LSTM-ATT	That sucks if you have to take the sats tomorrow.	Neutral ×
LSTM-VAT	That sucks if you have to take the sats tomorrow.	Negative √

Figure 7: Two examples of attention visualization. Red denotes the attentive weights of the words. A deeper color indicates a larger value.

	Dataset	0	1	2	3	4	5
LSTM-VAT	Twitter	70.75	71.62	70.96	70.67	71.06	70.98
	IMDB	88.16	88.98	88.22	88.84	88.14	88.60
BERT-VAT	Twitter	74.84	75.26	77.71	77.13	76.68	76.76
	IMDB	91.81	92.06	92.11	92.09	91.92	91.96

Table 6: The accuracy with different iteration number with our LSTM/BERT-VAT model.

Also, we draw change of the sentence representation with different iterations (Figure 6). Similarly, we observe that fine-tuning and compressing iteratively can improve the sentence representations. The samples with the same class are close, and the samples with different classes have a large distance.

5.5 Case Studies

To understand why our proposed VAT model is more effective than the standard attention-based model, we visualize two examples of LSTM-based models using attention heatmaps (Figure 7). **First**, the standard attention-based LSTM model focuses on the wrong words (e.g., "this", "work") even though it predicts the right sentiment while our VAT model finds the correct words (e.g., "admired", "lot"). It indicates integrating IB into attention can help it focus on the key words and reduce the noisy information. **Second**, our proposed model can also improve the attention's performance by capturing the critical words accurately. For example, in the sentence "That sucks if you have to take the sats tomorrow.", our model predicts the right class label by attending the words "sucks" and "have to."

6 Conclusions and Future Work

This paper proposes a VAT-based framework to improve the performance and interpretability of attentions via both fine-tuning and compressing. The experimental results on eight benchmark datasets for text classification verify the effectiveness of

our models within this framework. In addition, we apply the framework for sentiment detection, which further demonstrates the superiority in terms of interpretability. It is also interesting to find that training the models by fine-tuning and compressing iteratively is effective to improve the text representations. In the future, we will investigate the effectiveness of our proposed attention framework for other tasks and areas, such as machine translation and visual question answering.

Acknowledgement

The authors wish to thank the reviewers for their helpful comments and suggestions. This research is (partially) supported by NSFC (62076097), STCSM (18ZR1411500), the Fundamental Research Funds for the Central Universities. This research is also funded by the Science and Technology Commission of Shanghai Municipality (19511120200 & 20511105102). The computation is performed in ECNU Multifunctional Platform for Innovation (001). The corresponding authors are Yuanbin Wu and Liang He.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Seojin Bang, Pengtao Xie, Heewook Lee, Wei Wu, and Eric Xing. 2019. Explaining a black-box using deep variational information bottleneck approach. *arXiv preprint arXiv:1902.06918*.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312.

Hanjie Chen and Yangfeng Ji. 2020. Learning variational word masks to improve the interpretability of neural text classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4236–4251.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a deep and unified understanding of deep neural models in nlp. In *International conference on machine learning*, pages 2454–2463. PMLR.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Zhiying Jiang, Raphael Tang, Ji Xin, and Jimmy Lin. 2020. Inserting information bottleneck for attribution in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3850–3857.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. 2017. Structured attention networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Xiang Lisa Li and Jason Eisner. 2019. Specializing word embeddings (for parsing) by information bottleneck. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2744–2754.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1101–1111.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078.

Bo Pang and Lillian Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Marek Rei and Anders Søgaard. 2018. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 293–302.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. 2020. Restricting the flow: Information bottlenecks for attribution. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Naftali Tishby, Fernando C Pereira, and William Bialek. 1999. The information bottleneck method. In *Proceedings of the 37th annual Allerton Conference on Communication, Control, and Computing*, page 368–377.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20.

Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard Hovy. 2017. An interpretable knowledge transfer model for knowledge base completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 950–962.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

Andrey Zhmoginov, Ian Fischer, and Mark Sandler. 2019. Information-bottleneck approach to salient region discovery. *arXiv preprint arXiv:1907.09578*.