Learning Intrinsic Dimension via Information Bottleneck for Explainable Aspect-based Sentiment Analysis

Zhenxiao Cheng[1], **Jie Zhou**[1,*], **Wen Wu**[1], **Qin Chen**[1], **Liang He**[1]
[1] School of Computer Science and Technology, East China Normal University, Shanghai, China

Abstract

Gradient-based explanation methods are increasingly used to interpret neural models in natural language processing (NLP) due to their high fidelity. Such methods determine word-level importance using dimension-level gradient values through a norm function, often presuming equal significance for all gradient dimensions. However, in the context of Aspect-based Sentiment Analysis (ABSA), our preliminary research suggests that only specific dimensions are pertinent. To address this, we propose the Information Bottleneck-based Gradient (IBG) explanation framework for ABSA. This framework leverages an information bottleneck to refine word embeddings into a concise intrinsic dimension, maintaining essential features and omitting unrelated information. Comprehensive tests show that our IBG approach considerably improves both the models' performance and interpretability by identifying sentiment-aware features.

Keywords: Intrinsic dimension, Information Bottleneck, Explainable, Aspect-based Sentiment Analysis

1. Introduction

The domain of natural language processing (NLP) has witnessed the rise of neural models that offer remarkable capabilities. Yet, the intricacies of these models often remain cloaked in layers of complexity, raising questions about their interpretability (Danilevsky et al., 2020; Ribeiro et al., 2016; Lundberg and Lee, 2017). Gradient-based explanation methods (Simonyan et al., 2014; Smilkov et al., 2017; Sundararajan et al., 2017) have emerged as a prominent solution to this interpretability conundrum, offering insights into how neural models function (Doshi-Velez and Kim, 2017), especially in terms of their fidelity.

These methods pivot on the idea of ascertaining the importance of words by utilizing dimension-level gradient values, processed through a norm function. Formally, Gradient-based explanation methods estimate the contribution of input x towards output y by computing the partial derivative of y w.r.t x. These saliency methods can be used to enable feature importance explainability, especially on word/token-level features (Aubakirova and Bansal, 2016; Karlekar et al., 2018). Then, Smooth Gradient (Smilkov et al., 2017) and Integrated Gradients (Sundararajan et al., 2017) are proposed to improve the original gradients. A prevalent assumption made during this process is the uniform significance attributed to every gradient dimension.

However, while this might hold true for many applications, the nuances of Aspect-based Sentiment Analysis (ABSA) present a more complex scenario. In our preliminary analysis of aspect-based sentiment classification tasks (Section 2), we have found that this assumption is not always valid. **First**, not all dimensions are equally significant. **Second**, while the number of important dimensions varies across datasets, only a few dimensions prove essential (intrinsic dimension (Li et al., 2018)). **Third**, the key dimensions exhibit similarity within a dataset but vary across different datasets. For example, dimension 401 ranks among the top 100 important dimensions for 89% instances in the Res14 dataset. Further elaboration on these findings can be found in Section 2. This observation calls for a more discerning approach to gradient-based explanations in the ABSA context.

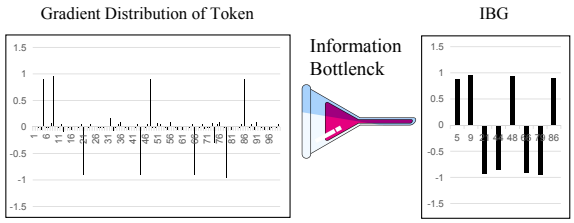
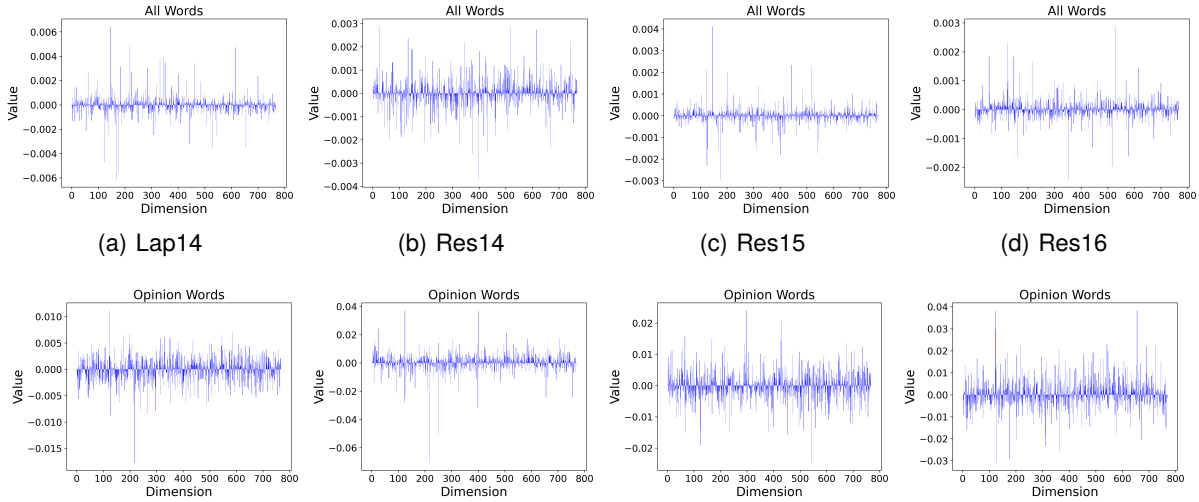


Figure 1: IBG compresses the 100 noisy dimensions of token gradient into 8 intrinsic dimensions via information bottleneck.

In this paper, we aim to answer the question: "How can important dimensions be dynamically selected?" We propose an Information Bottleneck-based Gradient explanation framework (IBG) for ABSA to learn the intrinsic dimension. To be specific, we propose an Information Bottleneck-based Intrinsic Learning (iBiL) structure to distill word embeddings into an intrinsic dimension that is both concise and replete with pertinent information, ensuring that irrelevant data is judiciously pruned (Figure 1). Our model is model-agnostic, we in-

* Corresponding author, jzhou@cs.ecnu.edu.cn.

LREC-COLING 2024, pages 10274–10285



(a) Lap14 (b) Res14 (c) Res15 (d) Res16

(e) Lap14 (f) Res14 (g) Res15 (h) Res16

Figure 2: Visualization of the gradients on hidden dimension over four classic ABSA datasets.

tegrate it with several state-of-the-art baselines, such as BERT-SPC (Kenton and Toutanova, 2019) and DualGCN (Li et al., 2021). Our comprehensive evaluations and tests provide substantial evidence of the effectiveness of the IBG framework. As detailed in the following sections, IBG not only enhances the performance metrics but also improves the clarity of interpretations, shedding light on sentiment-aware features.

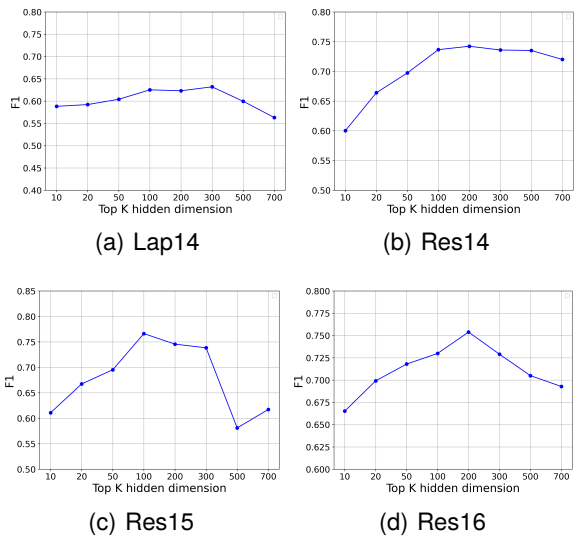
The key contributions of this paper are listed as follows[1].

- We propose the Information Bottleneck-based Gradient (IBG) explanation framework to find the low-dimensional intrinsic space since we discover that not all dimensions of the embedding are equally important in completing the ABSA task through preliminary analyses.
- We introduce the iBiL structure, forcing the model to learn its intrinsic sentiment embedding by effectively removing irrelevant information while retaining sentiment-related details via information bottleneck.
- Through extensive experiments, we demonstrate that our framework is capable of enhancing both the performance and the interpretability of the original model significantly.

2. Preliminary Analysis

In this section, we mainly conduct preliminary analysis to answer the following three questions.

[1]Our code is publicly available at https://github.com/sofistikate/IBG



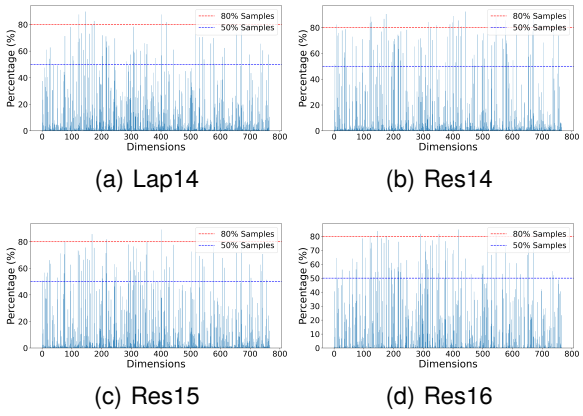
(a) Lap14 (b) Res14

(c) Res15 (d) Res16

Figure 3: The influence of top k hidden dimension

Question 1: Are all the hidden dimensions equally important? In Figure 2, we visualize the gradients of the hidden dimensions. Specifically, we train a sentiment classifier for aspect-based sentiment analysis, which inputs the sentence and aspect into a classifier to predict the sentiment polarity. Then we compute the gradients of the opinion words concerning the given aspects as well as all the words in the sentence. The classifier we used is the Bert-SPC model, which is a very classic and fundamental baseline.

Our observations from these figures can be summarized as follows: **First**, we note substantial variations in the values along the dimensions. Several gradient values are notably larger than others, indicating that only a small fraction of dimen-



(a) Lap14 (b) Res14

(c) Res15 (d) Res16

Figure 4: Percentage of Samples Where Each Dimension is in the Top 100 Importance

sions consistently contributes significantly to the model's predictions for each sample. Most dimensions have limited influence. **Second**, the values of the opinion words are considerably larger than those of general words. This suggests that gradients can assist in identifying key words.

Question 2: How many dimensions are necessary? The fact that all dimensions do not have equal importance implies that there exist certain dimensions that are really important. Therefore, in this subsection, we will discuss what is the exact number of key dimensions among all dimensions. Specifically, we employ gradient-based explanation methods to select the top-k important dimensions for predicting sentiment polarity with respect to a given aspect (see Figure 3).

Our observations reveal the following: **First**, utilizing approximately the top 100-300 dimensions often yields similar results to using all dimensions in most cases. **Second**, in certain instances, employing the top-k dimensions can lead to performance improvements compared to using all dimensions.

Question 3: Which dimensions are important? Therefore, aiming to ascertain whether certain dimensions consistently play a crucial role in ABSA, we conduct a statistical analysis focusing on the top-100 important dimension indices in each sample's prediction, as detailed in Figure 4.

Our findings indicate the following: **First**, it is noteworthy that across more than 50% of samples within each dataset, approximately 80 dimensions consistently maintain their positions within the top-100 important dimension indices. **Second**, several dimension indices consistently demonstrate significance in nearly every sample within a dataset. For instance, the top-3 significant dimensions are (145, 124, 401), (443, 175, 401), (401, 168, 218) and (419, 145, 219) in the Lap14, Res14, Res15 and Res16 datasets, respectively. However, it is essential to recognize that the specific important dimensions vary among different datasets.

3. Our Approach

In this paper, we propose the IBG explanation framework for ABSA by learning the low-dimensional intrinsic features via information bottleneck (Figure 5). Our framework explains the sentiment classifier by extracting the aspect-aware opinion words via the gradient method (Section 3.1). It calculates the important weights according to the gradients on the embedding level, which contains redundancy information. Thus, we introduce our self-designed Information Bottleneck-based Intrinsic Learning (iBiL) structure between the embedding layer and encoder layer of the traditional language model (Section 3.2). This incorporation utilizes the information bottleneck principle to compress the embedding layer to get the intrinsic representations, thereby eliminating redundant information in the original embedding and retaining essential information.

Let s be a sentence with words $\{w_1, w_2, ..., w_{|s|}\}$, a be an aspect in the sentence s, and $y \in Y$ be the sentiment label of a. Given a corpus $\mathcal{D} = \{(s_i, a_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, our task is to predict the sentiment polarity $y \in \{P, N, O\}$ of the sentence towards the given aspect a. P, N, O represent positive, negative and neutral, respectively. The word embeddings of the sentence s are $x = \{x_1, ..., x_i, ..., x_{|s|}\}$, where x_i is the word embedding of w_i. Moreover, we aim to explain the model by extracting the *aspect-aware opinion words* o that express the sentiment w.r.t the aspect a.

3.1. Overview of IBG

In this section, we introduce the overall structure of our IBG. It is a gradient-based explanation method based on an aspect-based sentiment classifier, which computes the importance scores for each token based on the embedding to find the aspect-specific opinion words (e.g., delicious) To improve the interpretability, it incorporates our designed iBiL structure between the embedding and encoder layers. This structure is employed to acquire intrinsic representations at the embedding level, eliminating redundant information while retaining emotion-related essential information.

The prevailing paradigm of gradient-based model interpretation methods in NLP consists of two main steps. First, a pre-trained language model, including the embedding layer, is fine-tuned for specific downstream tasks. Subsequently, importance scores for each token are computed, primarily through the embedding layer.

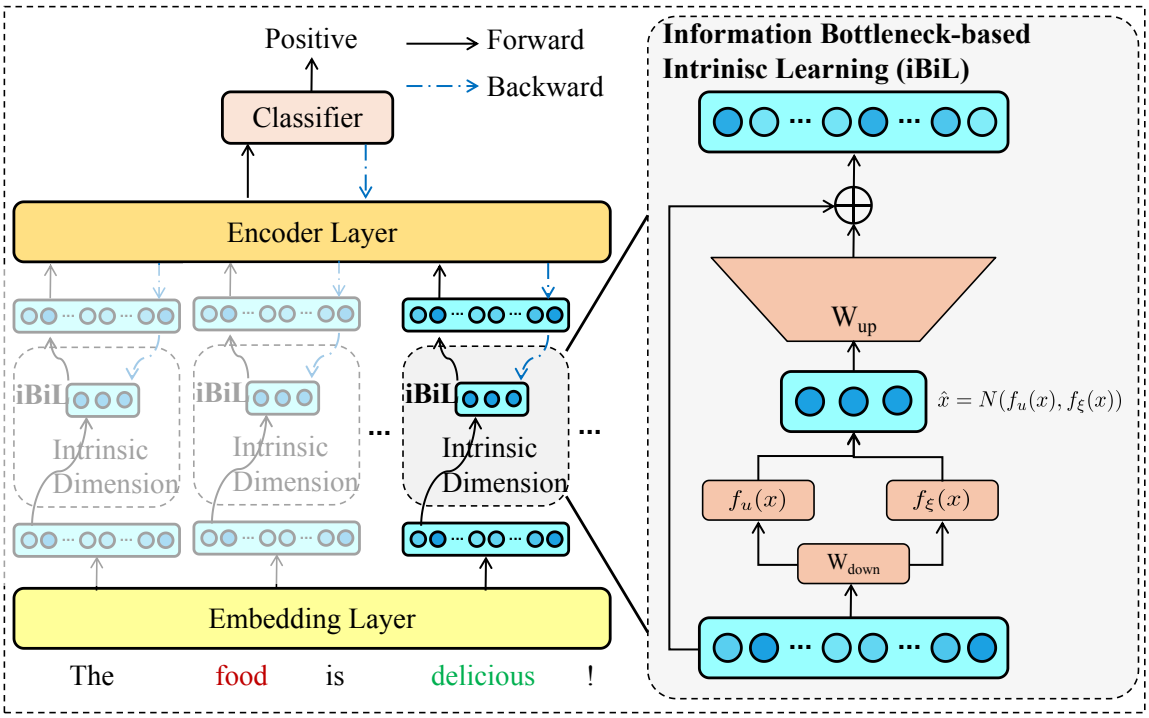


Figure 5: The framework of our IBG approach.

Particularly, we train a sentiment classifier using our iBiL for ABSA, which aims to predict the sentiment of the sentence concerning the given aspect. Let $\mathcal{F}$ be a sentiment classifier with embedding and encoder layers that predicts the sentiment distribution $P(y|s,a)$ based on the sentence s and aspect a.

$$P(y|s,a) = \mathcal{F}(s,a) \tag{1}$$

It is worth noting that, given our framework's model-agnostic nature, the sentiment classifier $\mathcal{F}$ can be any existing ABSA model. In this paper, we mainly conduct the experiments on BERT-SPC (Kenton and Toutanova, 2019) and DualGCN-BERT (Li et al., 2021).

After the words w in the sentence s passing through the embedding layer, we obtain embeddings $x = \{x_1, ..., x_i, ..., x_{|s|}\}$, where $x_i \in \mathcal{R}^{\text{High}}$ is the high dimension embedding of word w_i. Afterward, we employ iBiL structures to learn distinct intrinsic dimensional representations via the information bottleneck structure, which will be introduced in Section 3.2. Through iBiL, we also obtain the "intrinsic representation" $\hat{x} = \{\hat{x}_1, ..., \hat{x}_i, ..., \hat{x}_{|s|}\}$, where $\hat{x}_i \in \mathcal{R}^{\text{Low}}$ is the low dimension representation of word w_i. Here, High is the dimensionality of the hidden layer in the original embedding, while Low represents the size of the intrinsic dimensionality, where High (e.g., 768) is much larger than Low (e.g., 10, 20). To input the final word representation x' into the encoder layer, we upsample the intrinsic representation into a hidden dimension.

$$x' = x + W_{up}\hat{x} \tag{2}$$

where W_{up} represents a learnable upsampling matrix. Throughout the entire training process, x remains frozen and unchanged.

Subsequently, a novel attribution method is employed to determine the importance scores for each token based on both the original embeddings and intrinsic representations, ultimately leading to improvements in models' performance and interpretability. Similar to prior work (Simonyan et al., 2014), we compute importance scores by dotting the gradients on the model's prediction with the corresponding vectors:

$$\begin{aligned} \gamma(w_i) &= \left| x_i \times \frac{\partial \mathcal{F}(s,a)}{\partial x_i} \right| \\ \hat{\gamma}(w_i) &= \left| \hat{x}_i \times \frac{\partial \mathcal{F}(s,a)}{\partial \hat{x}_i} \right| \end{aligned} \tag{3}$$

where γ and $\hat{\gamma}$ are the importance weights for each token in the embedding layer and the intrinsic representation layer.

Finally, for each token's γ and $\hat{\gamma}$, we introduce a hyperparameter α to balance the weights between the two scores. For each word w_i, the final importance score is calculated as:

$$\text{FScore}(w_i) = (1-\alpha)\gamma(w_i) + \alpha\hat{\gamma}(w_i) \tag{4}$$

Finally, we use $\text{FScore}(w_i)$ to select the opinion words of aspect.

3.2. Information Bottleneck-based Intrinsic Learning

In our proposed IBG framework, one of the key innovations lies in the design of Information Bottleneck-based Intrinsic Learning (iBiL). This involves compressing the pre-trained embedding x into a lower-dimensional space $\hat{x}$ to learn the intrinsic sentiment representation in the embedding layer. The primary goal is to retain essential information in the embedding while removing redundant noise.

According to the original IB theory (Alemi et al., 2016), the main objective is to learn a compressed representation Z by maximizing the mutual information between Z and output Y and minimizing the mutual information between Z and input X. In our work, to learn the intrinsic representation $\hat{x}$, we aim to preserve the information of $\hat{x}$ with respect to the sentiment polarity y while minimizing the mutual information between $\hat{x}$ and the original word vectors x. Therefore, our goal is to minimize the following loss:

$$\mathcal{L} = \beta I(x; \hat{x}) - I(y; \hat{x}) \tag{5}$$

where $I(.,.)$ represents the mutual information.

However, in practical operations, directly calculating the mutual information x and $\hat{x}$ is not feasible. We employ the Variational Inference method (Li and Eisner, 2019) to estimate the final loss function. Specifically, the upper bound of $I(x; \hat{x})$ can be obtained through the following method,

$$I(x; \hat{x}) \leqslant \int d\hat{x}dx p(x)p(x, \hat{x}) \log \frac{p(\hat{x} \mid x)}{q(\hat{x})} \tag{6}$$

Since

$$\begin{aligned} &KL[p(\hat{x}), q(\hat{x})] \geqslant 0 \\ \Longrightarrow &\int d\hat{x} p(\hat{x}) \log p(\hat{x}) \geqslant \int d\hat{x} p(\hat{x}) \log q(\hat{x}) \end{aligned} \tag{7}$$

In actual computation process, we first sample $\hat{x}$ from the original word vectors using the reparameterization method (Kingma and Welling, 2014):

$$p(\hat{x} \mid x) = f_\mu(x) + f_\xi(x) \cdot z \tag{8}$$

where $z \sim \mathcal{N}(0, I)$. Here, f_μ and f_ξ correspond to two linear layers, which are trainable. And $q(\hat{x})$ in Equation 6 is assumed to follow the standard normal distribution. Therefore, $I(x; \hat{x})$ in Equation 5 can be directly replaced with the KL loss in Equation 6 and be easily calculated.

The lower bound of $I(y; \hat{x})$ can also be obtained as follows,

$$I(y; \hat{x}) \geqslant \int dyd\hat{x} p(y, \hat{x}) \log q(y \mid \hat{x}) - \int dy p(y) \log p(y) \tag{9}$$

Since

$$\begin{aligned} &KL[p(y \mid \hat{x}), q(y \mid \hat{x})] \geqslant 0 \implies \\ &\int dy p(y \mid \hat{x}) \log p(y \mid \hat{x}) \geqslant \int dy p(y \mid \hat{x}) \log q(y \mid \hat{x}) \end{aligned} \tag{10}$$

The $\int dyp(y) \log p(y)$ in Equation 9 is a constant, and thus, we do not need to pay attention to it during the optimization. If we consider $q(y \mid \hat{x})$ as the subsequent dimensionality-expanding matrix and encoder structure of the model, then the process of minimizing $I(y; \hat{x})$ can be directly approximated as optimizing the original sentiment classification loss function. In other words, $I(y; \hat{x})$ can be straightforwardly replaced by a cross-entropy loss $\mathcal{L}_{CE}$. The final loss function can be written as follows.

$$\mathcal{L} = \mathcal{L}_{CE} + \beta KL[p(\hat{x} \mid x), q(\hat{x})] \tag{11}$$

4. Experiment Setups

4.1. Datasets, Metrics and Settings

Datasets. To evaluate the performance and the interpretability of our IBG framework, we conduct a comprehensive series of experiments on four common datasets: Res14, Lap14, Res15 and Res16 (Fan et al., 2019), which labeled the opinion words for each aspect.

Metrics. In order to assess the performance of our framework, we employ the most widely used metrics in the ABSA task: Accuracy and Macro F1-score. Moreover, to verify the effectiveness of our framework in improving models' interpretability, following (Chen and Ji, 2020), we adopt the area over the perturbation curve (AOPC) (Nguyen, 2018; Samek et al., 2016) and Post-hoc Accuracy (Ph-Acc) (Chen et al., 2018). AOPC computes the average decrease in accuracy when the model makes predictions after removing the top-k important words for explanation. Ph-acc, on the other hand, retains only the top-k words and masks the remaining words to assess whether the model can still make accurate predictions.

Settings. We use bert-base-uncased version (Kenton and Toutanova, 2019) and Adam optimizer for the original BERT and the additional components we introduce with learning rates 1e-5 and 1e-4, respectively. In the selection of intrinsic dimensionality sizes, we consider dimensions of 5, 10, 20, 50, 100 and 300. The hyperparameter α was set to 0.5.

4.2. Baselines

We compare our framework with the SOTA baselines to investigate its performance and inter-

Model	Lap14				Res14				Res15				Res16			
	Acc	F1	AOPC	Ph-Acc	Acc	F1	AOPC	Ph-Acc	Acc	F1	AOPC	Ph-Acc	Acc	F1	AOPC	Ph-Acc
AEN-BERT	81.80	56.07	-	-	88.59	72.69	-	-	86.44	63.73	-	-	88.60	65.06	-	-
LCF-BERT	81.83	58.23	-	-	90.00	67.91	-	-	85.94	67.53	-	-	89.91	69.98	-	-
ASCM4ABSA	81.93	57.34	-	-	89.96	70.85	-	-	86.81	66.32	-	-	88.98	68.04	-	-
ChatGPT	82.78	44.60	-	-	90.62	51.72	-	-	90.54	71.86	-	-	93.63	76.70	-	-
BERT-SPC$_{\texttt{IEGA}}$	82.28	62.93	15.04	42.18	90.62	72.75	11.13	69.18	85.40	59.39	08.26	70.78	88.56	62.60	10.48	75.49
RGAT-BERT$_{\texttt{IEGA}}$	82.58	65.10	13.58	66.52	91.64	77.50	15.67	63.29	87.09	69.36	16.07	72.98	90.78	67.34	12.71	80.04
BERT-SPC$_{\texttt{Grad}}$	81.80	56.31	11.13	67.02	90.47	72.01	09.88	62.71	88.02	61.73	10.60	76.73	90.45	70.32	10.53	79.17
BERT-SPC$_{\texttt{InteGrad}}$	-	-	09.85	74.09	-	-	10.47	54.24	-	-	16.59	80.88	-	-	10.31	81.40
BERT-SPC$_{\texttt{SmoothGrad}}$	-	-	09.21	69.81	-	-	12.59	69.65	-	-	12.67	78.11	-	-	09.43	81.58
DualGCN-BERT$_{\texttt{Grad}}$	84.27	67.07	18.84	64.60	91.41	76.62	11.65	77.65	88.71	69.82	13.36	73.04	91.67	76.97	13.60	73.46
DualGCN-BERT$_{\texttt{InteGrad}}$	-	-	14.35	59.46	-	-	13.06	74.35	-	-	14.29	74.19	-	-	14.91	71.71
DualGCN-BERT$_{\texttt{SmoothGrad}}$	-	-	22.70	68.03	-	-	15.53	80.24	-	-	14.29	74.19	-	-	16.23	75.00
BERT-SPC$_{\texttt{IBG}}$	83.30	65.34	17.77	**75.16**	91.29	76.33	**28.35**	**84.24**	89.40	**78.32**	17.28	**81.34**	92.32	79.57	20.39	**83.33**
DualGCN-BERT$_{\texttt{IBG}}$	**85.22**	**72.99**	**26.89**	66.17	**92.58**	**80.22**	27.18	78.71	**90.63**	77.68	**25.58**	75.81	**93.64**	**84.61**	**20.83**	82.24

Table 1: The main results of performance and interpretability.

pretability. To validate the performance, we select the following baselines:

- BERT-SPC (Kenton and Toutanova, 2019) simply concatenates the raw sentences with the corresponding aspect terms, subsequently feeding these inputs directly into a pre-trained BERT model for ABSA.
- AEN-BERT (Song et al., 2019) proposes an Attentional Encoder Network (AEN) and enhances BERT with attention-based encoders to capture context-specific information related to aspects.
- LCF-BERT (Zeng et al., 2019) designs a Local Context Focus (LCF) mechanism which uses multi-head self-attention to force the model to pay attention to the local context words.
- RGAT-BERT (Wang et al., 2020) is the first work that uses the Graph Convolutional Network (GCN) in ABSA to utilize the syntactical dependency structures in the sentences.
- DualGCN-BERT (Li et al., 2021) is a Dual GCN, employing two GCNs, which can simultaneously capture both syntax and semantic information.
- ASCM4ABSA (Ma et al., 2022) proposes three aspect-specific input methods and exploits these transformations to promote the language models to pay more attention to the aspect-specific context in ABSA.
- ChatGPT (Wang et al., 2023) designs a prompt specifically for ABSA, using the GPT-3.5-turbo model to generate the results and analyze its performance.
- IEGA (Cheng et al., 2023) is a model agnostic Interpretation-Enhanced Gradient-based framework for ABSA, which guides the model's attention towards important words like opinion words.

In addition, to verify the interpretability, we select some classic gradient-based explanation strategies for comparison.

- Simple Gradient (Simonyan et al., 2014) calculates the gradient of the model's output w.r.t. the input by taking the dot product of these gradients with the corresponding feature values.
- Smooth Gradient (Smilkov et al., 2017) is an enhancement of the Simple Gradient that reduces noise and provides more stable explanations by computing the gradient at multiple points along the path to the actual input.
- Integrated Gradients (Sundararajan et al., 2017) considers a baseline input and calculates the gradient at multiple points along the straight-line path, connecting the baseline to the actual input.

5. Results and Analyses

5.1. Main Results

To demonstrate that our framework can both enhance models' performance and interpretability, we conducted extensive comparative experiments (Table 1). The results yield the following findings:

First, IBG outperforms other models in terms of both performance and interpretability. Regarding performance, our framework performs better than the strong models like RGAT and DualGCN, as well as the latest ChatGPT model. As for interpretability, IBG also better captures meaningful words compared to classical gradient-based strategies and model-agnostic methods like IEGA. This indicates that IBG can effectively capture the sentiment-related low-dimensional features.

Second, IBG is capable of further enhancing both the performance and interpretability based on the existing models. We can see that our IBG results in a 1-2 points improvement in accuracy on each dataset. Taking BERT-SPC as an example, after integrating our framework, the Macro-F1 is improved by more than 10 points on the Lap14, Res15 and Res16 datasets. Even though Dual-GCN already performs well, Dual-GCN$_{\texttt{IBG}}$

Model	Lap14				Res14				Res15				Res16			
	Acc	F1	AOPC	Ph-Acc	Acc	F1	AOPC	Ph-Acc	Acc	F1	AOPC	Ph-Acc	Acc	F1	AOPC	Ph-Acc
BERT-SPC$_{\texttt{IBG}}$	83.30	65.34	17.77	75.16	91.29	76.33	28.35	84.24	89.40	78.32	17.28	81.34	92.32	79.57	20.39	83.33
w/o IB	82.87	64.66	18.63	71.09	90.94	74.62	28.94	72.71	89.40	76.14	17.28	78.57	91.45	78.54	17.76	82.46
w/o iBiL	81.80	56.31	11.13	67.02	90.47	72.01	09.88	62.71	88.02	61.73	10.60	76.73	90.45	70.32	10.53	79.17

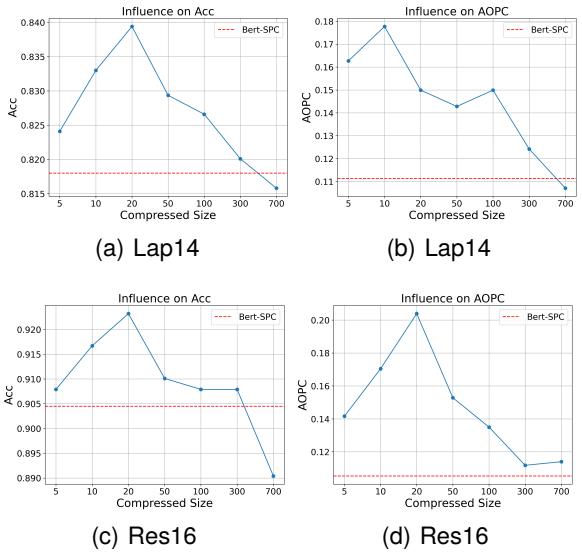
Table 2: The results of ablation studies.

still achieves an approximately 1% increase in accuracy and 5% improvement in F1 score over all datasets. We also notice that our framework exhibits a higher improvement in Macro-F1. This indicates that the representations obtained by information bottleneck demonstrate higher sensitivity to different polarities, including neutral sentiment. On the other hand, the effect of IBG on enhancing model interpretability is also significant. The models exhibit substantial increases in both AOPC and Ph-Acc after the incorporation of IBG. Compared to IEGA, the AOPC of BERT-SPC$_{\texttt{IBG}}$ surpasses BERT-SPC$_{\texttt{IEGA}}$ by 17.22%, 9.02% and 9.91% on the Res14, Res15 and Res16 datasets, respectively. These results further verify our IBG's ability to better capture keywords like opinion words through learning the intrinsic dimension.

Third, IBG offers a better explanation in comparison to traditional gradient-based model interpretation strategies (Simple Gradient, Smooth Gradient and Integrated Gradient). BERT-SPC with IBG demonstrates more than twofold improvement in AOPC on the Res14 and Res16 datasets. Ph-Acc shows a remarkable 40.00% improvement on Res14 compared with the Integrated Gradient method. Similarly, DualGCN-BERT$_{\texttt{IBG}}$ exhibits noticeable enhancements in AOPC compared to these three interpretation strategies, with the greatest improvements of 12.54%, 15.53%, 12.22% and 7.23% on the four datasets. However, the improvement in Ph-Acc on DualGCN-BERT is less pronounced. This can be attributed to the fact that DualGCN leverages GCN to capture the dependency relationships among words. Consequently, masking excessive words within a sentence can disrupt the integrity of the graph's structure, thus impacting the final prediction. In summary, the incorporation of our framework is capable of enhancing models' ability to focus on contextually implicit sentiment information in words through gradient analysis, thereby improving its explanation.

5.2. Ablation Studies

We also conduct ablation experiments (Table 2) to validate our framework from two perspectives: First, removing the iBiL (w/o iBiL) reduces performance, which indicates that compressing the model into a low-dimensional space for learning intrinsic dimension does benefit the ABSA task. Second, the information bottleneck module has a significant impact on the model's final performance and interpretability. After removing the information bottleneck structure (w/o IB), the model's accuracy and F1-score both decline, even with a decrease of 4.07% and 11.53% in Ph-Acc on the Lap14 and Res14 datasets, respectively. This indicates that simply compressing embeddings to a lower dimensionality introduces noise, which affects the model's classification ability. Introducing an information bottleneck, on the other hand, does enable the model to forget irrelevant information and retain important features for sentiment classification.



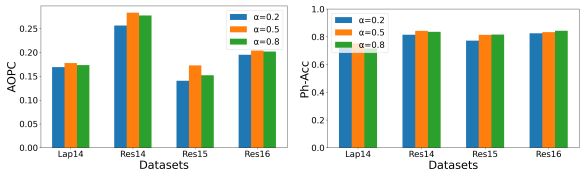
(a) Lap14 (b) Lap14

(c) Res16 (d) Res16

Figure 6: The influence of compressed size Low.

5.3. Further Analysis

The Influence of Compressed Size Low**.** We further explore the variation in model performance and interpretability when compressing the pretrained BERT embeddings into different dimensions. Two key findings can be deduced from Figure 6: First, with the continuous increase in dimensions, both models' performance and interpretability exhibit a trend of initially increasing and then decreasing. Second, the optimal state of the model consistently reaches at 10 or 20 dimensions and the performance is even better than the original model without compressing. This suggests that our framework can effectively capture essential information in models' embeddings and eliminate redundant information.



(a) AOPC (b) Ph-Acc

Figure 7: The influence of α on four datasets

The Influence of α. We conduct experiments to observe the impact of varying α values in Equation 4, which are used to balance the weights between the important scores obtained from intrinsic features and the original word embeddings (Figure 7). We can see that both the intrinsic representations and the original word embeddings are useful for the model's explanation. Relatively, assigning a higher weight to intrinsic dimension results in slightly higher AOPC and Ph-Acc compared with assigning a higher weight to word embeddings. It shows intrinsic vectors successfully learn the aspect-aware sentiment features, which is important for ABSA.

6. Related Work

6.1. Intrinsic Dimension

Oymak et al. (2021) showed that trained neural networks often exhibit a low-rank property, giving rise to the concept of "intrinsic dimension" (Li et al., 2018), which represents the minimum subspace dimension capable of encoding effective information or solving a problem. Following this concept, Pope et al. (2020) applied dimension estimation techniques to high-dimensional image data and discovered that natural image datasets have low intrinsic dimension. Significantly, Aghajanyan et al. (2021) first introduced the intrinsic dimension into the field of natural language processing, revealing that the size of the large pre-trained language models' intrinsic dimension is much smaller compared to that of their whole parameters. After this, many studies started to incorporate low-rank structures into large models, achieving parameter-efficient learning by only tuning a small subspace based on the intrinsic dimension Qin et al. (2021); Sun et al. (2022b,a); Houlsby et al. (2019); Wang et al. (2022); Hu et al. (2021). Differing from predecessors' attempts to compress or tune models using intrinsic dimension, our work learns the intrinsic dimension of the large language models via information bottleneck, exploring how to better interpret the model's predictions in ABSA.

6.2. Gradient-based Explanation Algorithms

Literature on explanation and attribution methods has grown in the last few years, with a few broad categories of approaches: perturbing the input (Fong et al., 2019; Ribeiro et al., 2016); utilizing gradient (Baehrens et al., 2010; Binder et al., 2016; Selvaraju et al., 2017); visualizing intermediate layers (Zeiler and Fergus, 2014). Our work extends and improves upon the gradient-based method (Simonyan et al., 2014), a popular technique applicable to many different types of models. Several works were proposed to improve the original gradient-based methods, such as SmoothGrad (Smilkov et al., 2017) and Integrated Gradient (Sundararajan et al., 2017). Different from them, we optimize to compute gradient scores based on an intrinsic space, enabling a more effective model interpretation method.

6.3. Information Bottleneck

A series of studies motivated us to utilize IB (Li and Eisner, 2019; Zhou et al., 2021, 2022) to improve the explanations of gradient-based explanation methods. 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. (2021) 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., 2021) 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. In this paper, we utilize IB to learn the intrinsic space to improve the models' explanation.

6.4. Aspect-based Sentiment Analysis

Aspect-based Sentiment Analysis (ABSA) involves the extraction of aspect terms and opinion words from a sentence and the prediction of sentiment polarity (Zhang et al., 2022a). In our study, we focus on the subtask Aspect-based Sentiment Classification (ABSC), which entails predicting sentiment labels for a given sentence and its associated aspect. To consider the complex contextual relationships in sentences, some ABSC research combined attention mechanisms with large pre-trained language models, such as BERT-SPC (Kenton and Toutanova, 2019), AEN-BERT (Song et al., 2019) and LCF-BERT (Zeng et al., 2019). There is another trend of combining dependency trees and Graph Convolutional Networks (GCNs), exploiting syntax information explicitly, like RGAT (Wang et al., 2020), DualGCN (Li et al., 2021) and

SSEGCN (Zhang et al., 2022b). However, even when employing methods such as dependency trees to align aspect terms with their corresponding opinion words, in practice, we still observe that the model may focus on the wrong aspect, especially in sentences containing multiple aspects. Therefore, research on Explainable Aspect-based Sentiment Analysis is essential. Cheng et al. (2023) leveraged annotated opinion words to force the model to pay greater attention to these words in terms of gradients, enhancing the models' interpretability, but the cost of annotation is high. In this paper, we propose a model-agnostic framework to enhance both performance and interpretability without additional labels.

7. Conclusions and Further Work

This paper conducts preliminary experiments, demonstrating the uneven importance of word embedding dimensions in ABSA. However, the current gradient-based explanation methods do not take this difference into account. Thus, we propose an Information Bottleneck-based Gradient (`IBG`) explanation framework for ABSA, leveraging the information bottleneck principle to compel the model to learn intrinsic information. By integrating our framework with the latest models, we conduct extensive comparative experiments, confirming that our proposed `IBG` framework significantly enhances both the performance and interpretability of the original models. Through ablation experiments, we demonstrate the beneficial impact of the information bottleneck structure and the attempt to map the embedding layer to a low-dimensional intrinsic space. Future research will explore applying `IBG` to large-scale language models (e.g., LLaMA) and other NLP tasks.

Acknowledge

The authors wish to thank the reviewers for their helpful comments and suggestions. This research is funded by the National Key Research and Development Program of China (No.2021ZD0114002), the National Natural Science Foundation of China (No.62307028 and No.62377013), the Science and Technology Commission of Shanghai Municipality Grant (No.22511105901, No.21511100402 and No.21511100302), and Shanghai Science and Technology Innovation Action Plan (No.23ZR1441800 and No.23YF1426100).

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 7319–7328.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. In *Proceedings of the 4th International Conference on Learning Representations*.

Malika Aubakirova and Mohit Bansal. 2016. Interpreting neural networks to improve politeness comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2035–2041.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831.

Seojin Bang, Pengtao Xie, Heewook Lee, Wei Wu, and Eric Xing. 2021. Explaining a black-box by using a deep variational information bottleneck approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11396–11404.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Proceedings of the 25th International Conference on Artificial Neural Networks*, pages 63–71.

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*, pages 4236–4251.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 883–892.

Zhenxiao Cheng, Jie Zhou, Wen Wu, Qin Chen, and Liang He. 2023. Tell model where to attend: Improving interpretability of aspect-based sentiment classification via small explanation annotations. In *Proceedings of the 2023 International*

Conference on Acoustics, Speech and Signal Processing, pages 1–5.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 2509–2518.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2950–2958.

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 *Proceedings of the 36th International Conference on Machine Learning*, pages 2454–2463.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *Proceedings of the 9th International Conference on Learning Representations*.

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.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 701–707.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. In *Proceedings of the 6th International Conference on Learning Representations*.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6319–6329.

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*, pages 2744–2754.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.

Fang Ma, Chen Zhang, Bo Zhang, and Dawei Song. 2022. Aspect-specific context modeling for aspect-based sentiment analysis. In *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*, pages 513–526.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 1069–1078.

Samet Oymak, Mingchen Li, and Mahdi Soltanolkotabi. 2021. Generalization guarantees for neural architecture search with

train-validation split. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8291–8301.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. 2020. The intrinsic dimension of images and its impact on learning. In *Proceedings of the 9th International Conference on Learning Representations*.

Yujia Qin, Xiaozhi Wang, Yusheng Su, Yankai Lin, Ning Ding, Jing Yi, Weize Chen, Zhiyuan Liu, Juanzi Li, Lei Hou, et al. 2021. Exploring universal intrinsic task subspace via prompt tuning. *arXiv preprint arXiv:2110.07867*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

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 TNNLS*, 28(11):2660–2673.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. 2020. Restricting the flow: Information bottlenecks for attribution. In *Proceedings of the 8th International Conference on Learning Representations*.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of 2nd International Conference on Learning Representations*.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. In *Proceedings of the International Conference on Machine Learning Workshop*.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Tianxiang Sun, Zhengfu He, Hong Qian, Yunhua Zhou, Xuan-Jing Huang, and Xipeng Qiu. 2022a. Bbtv2: towards a gradient-free future with large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3916–3930.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022b. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on Machine Learning*, pages 20841–20855.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan, and Jianfeng Gao. 2022. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5744–5760.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of the 13th European Conference on Computer Vision*, pages 818–833.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16):3389.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022a. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

Zheng Zhang, Zili Zhou, and Yanna Wang. 2022b. Ssegcn: Syntactic and semantic enhanced graph convolutional network for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4916–4925.

Andrey Zhmoginov, Ian Fischer, and Mark Sandler. 2021. Information-bottleneck approach to salient region discovery. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference*, pages 531–546.

Jie Zhou, Yuanbin Wu, Qin Chen, Xuan-Jing Huang, and Liang He. 2021. Attending via both fine-tuning and compressing. In *Findings of the Association for Computational Linguistics*, pages 2152–2161.

Jie Zhou, Qi Zhang, Qin Chen, Liang He, and Xuan-Jing Huang. 2022. A multi-format transfer learning model for event argument extraction via variational information bottleneck. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1990–2000.