

Interlock-Free Multi-Aspect Rationalization for Text Classification

Shuangqi Li and Diego Antognini and Boi Faltings

École Polytechnique Fédérale de Lausanne, Switzerland

firstname.lastname@epfl.ch

Abstract

Explanation is important for text classification tasks. One prevalent type of explanation is rationales, which are text snippets of input text that suffice to yield the prediction and are meaningful to humans. A lot of research on rationalization has been based on the selective rationalization framework, which has recently been shown to be problematic due to the interlocking dynamics (Yu et al., 2021). In this paper, we show that we address the interlocking problem in the multi-aspect setting, where we aim to generate multiple rationales for multiple outputs. More specifically, we propose a multi-stage training method incorporating an additional self-supervised contrastive loss that helps to generate more semantically diverse rationales. Empirical results on the beer review dataset show that our method improves significantly the rationalization performance.

1 Introduction

Text classification is a common application of deep neural models (Kim, 2014; Conneau et al., 2017). However, lack of interpretability of the predictions is preventing deep models from being applied in critical fields. A prevalent way of explaining the predictions of text classification is selective rationalization. The key idea is to select informative text snippets of the input texts. If they are short and coherent enough to be understood by humans and suffice to yield the prediction as a substitute of the full text, they are called rationales (Lei et al., 2016).

A line of research has focused on models that are inherently interpretable, i.e, able to produce the prediction along with rationale(s) or mask(s). Lei et al. (2016) proposed the first selective rationalization models that extract one chunk of text as an overall rationale to explain the prediction. Many works (Chang et al., 2019a; Yu et al., 2019; Chang et al., 2020; Antognini and Faltings, 2021; Antognini et al., 2021) follow this framework. Useful explanation can also be multi-aspected (An-

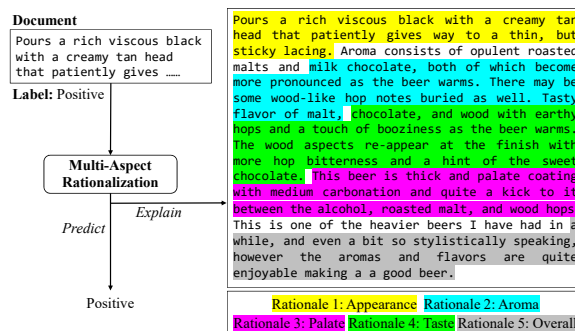


Figure 1: An illustration of multi-aspect rationalization. Given a beer review, the model generates five text snippets (i.e. rationales) that relate to different aspects of the beer, from which the prediction is computed. Different aspects are highlighted in different colors.

tognini et al., 2021; Antognini and Faltings, 2021), where each aspect is related to a particular concept, as illustrated in Figure 1. Unlike training multiple single-aspect rationale models in order to explain multiple outputs, one can train a single multi-aspect model. A significant advantage is that it only requires the overall label instead of labels for all aspects. This makes multi-aspect rationalization more practical.

Many works rely on the selective rationalization framework (Lei et al., 2016) that consists of a generator and a predictor or its variants. Intuitively, the generator extracts a snippet of text from the input and feeds it to the predictor to yield the classification. Training is essentially maximizing the mutual information between the selected text and the label. However, Yu et al. (2021) reveal the **interlocking** problem of this framework: the generator and predictor may get stuck in a suboptimal equilibrium. The interlocking dynamics prevents the generator from selecting the most informative text, and also prevents the predictor from seeing and predicting based on the most informative text.

In this work, we propose a new multi-stage training method that avoids the interlocking problem.

The method optimizes different objectives in three stages, incorporating a new self-contrastive loss function, which also promotes more semantically meaningful rationales. Experiments on the beer review dataset show that our multi-stage training fixes the interlocking problem and improves significantly the rationalization performance. Moreover, in a fully unsupervised setting, we show that the generator can learn even better rationalization using only the self-supervised contrastive loss.

2 Related Work

Selective rationalization (Lei et al., 2016) proposes the generator-predictor framework for rationalization. The generator can select rationales in a soft or hard way. Many works (Yu et al., 2019; Chang et al., 2020; Antognini and Faltings, 2021) use a hard constraint, forcing the generator to select text with a pre-specified length. Lei et al. (2016), Bastings et al. (2019) also propose to use a soft constraint to specify the sparsity level instead of the length. Antognini et al. (2021) propose to use a soft probabilistic mask and enable a more flexible rationalization with specified continuity level and sparsity level. The problems in the selective rationalization have raised attention. Chang et al. (2020) show that maximizing the mutual information can be problematic because it may pick up spurious correlations between input features and the output. Chang et al. (2019b) propose a game theoretic approach that captures the multi-faceted nature of rationales. Yu et al. (2021) reveal a major problem with the selective rationalization framework that impedes its performance on both classification and rationalization - model interlocking.

Multi-aspect rationalization Antognini et al. (2021) proposed the first multi-aspect rationalization model, MTM, which uses a probabilistic multi-

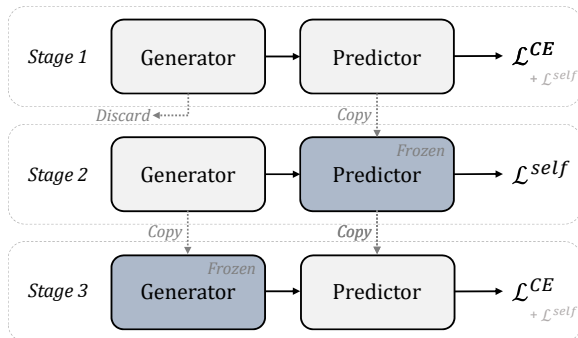


Figure 2: Our proposed interlock-free three-stage training method. Darker modules are frozen.

dimensional mask (one for each aspect) to explain multiple targets (one label for each aspect), and achieves higher classification accuracy and rationalization F1-score at the same time, even in the case of highly correlated data. Antognini and Faltings (2021) proposed a more practical multi-aspect model, ConRAT, that only requires the overall label. ConRAT identifies a set of concepts (candidate rationales) in the document and builds a selector to decide which ones are chosen, then aggregates the selected concepts to predict a single target variable.

3 Method

3.1 Multi-Aspect Rationalization

Let x denote the input text, composed of L words $(x^0, x^1, \dots, x^{L-1})$. The ground truth is a binary label y , telling the overall sentiment (positive or negative). K is predefined as the number of rationales to generate, which is equal to the number of aspects. The architecture is composed of two parts: 1) a rationale **generator** $g(\cdot)$ that takes x as input and outputs K rationales; 2) a **predictor** $f(\cdot)$ composed by a shared encoder followed by K binary classifiers. The shared encoder produces a representation for each selected rationale per aspect, the classifiers give a prediction for each aspect, which are then linearly aggregated into the final outcome.

To be more specific about the generator, Antognini et al. (2021) provide variable-length and soft rationales but use as many labels as rationales, while Antognini and Faltings (2021) leverages only the single overall label but is limited to hard rationales with strict continuity and predefined length. However, our model uses a soft generator to enable a more flexible rationalization, and we leverage only the overall label since the situation without aspect-wise labels is more realistic and common. In other words, we take the best of both worlds.

3.2 The Interlocking Dynamics

Yu et al. (2021) reveal a major problem with the selective rationalization framework - model interlocking. Ideally, the best generator $g(\cdot)$ and predictor $f(\cdot)$ can both be reached at the same time during the training. The predictor $f(\cdot)$ may overfit to the suboptimal rationales generated by the generator $g(\cdot)$ and keep reinforcing the generator's suboptimal behaviour. More formally, $\mathcal{L}^{CE}(g, f^*)$, where $f^* = \arg \min_f \mathcal{L}^{CE}(g, f)$ is **concave** with regard to g . Intuitively, the predictor can only see what the generator selects and

tends to overfit to the selection. Consequently, the predictor may produce a higher cross entropy loss even when the generator selects a better rationale than the current suboptimal one because the predictor has never seen the better rationale. Then the generator and predictor may get stuck in the suboptimal equilibrium. Yu et al. (2021) also proposed a framework called *A2R* that combines the selective rationalization paradigm and the attention-based explanation paradigm, where the concavity in the selective rationalization is mitigated or canceled by the convexity in the attention-based explanation. However, *A2R* is for single-aspect rationalization and cannot guarantee that the interlocking problem is completely avoided. Finally, it requires tuning a parameter to control the extent of added convexity.

3.3 Self-Supervised Contrastive Loss

One important feature in multi-aspect rationalization is to have diverse and discriminative rationales. We propose to use contrastive loss in our multi-stage training. Instead of the unsupervised contrastive loss applied in SimCLR (Chen et al., 2020) or MoCo (He et al., 2019), our self-supervised contrastive loss is more similar to the supervised contrastive loss (Khosla et al., 2020). The unsupervised contrastive loss contrasts an augmented version of each anchor sample against all other samples, regardless of the unavailable true labels, while the supervised contrastive loss is applied in the fully-supervised setting, leveraging the label information. It contrasts a set of samples from the same class against all other samples from different classes. Formally, the supervised contrastive loss (Khosla et al., 2020) for a batch of size N is

$$\begin{aligned} \mathcal{L}^{sup} &= - \sum_{i=1}^N \frac{1}{N_{y_i}} \sum_{j=1}^N \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{y_i = y_j} \cdot \mathcal{L}_{ij}^{sup} \\ \mathcal{L}_{ij}^{sup} &= \log \frac{\exp(z_i \cdot z_j / \tau_c)}{\sum_{l=1}^N \mathbb{1}_{i \neq l} \cdot \exp(z_i \cdot z_l / \tau_c)} \end{aligned} \quad (1)$$

where N_{y_i} denotes the number of samples that have the same label as the i^{th} sample in the batch, z_i denotes the representation for the i -th sample in the mini-batch, τ_c is the temperature hyperparameter.

In our multi-aspect setting, we can consider each rationale per aspect as a sample, and the generator gives K samples (rationales) for each input text. The label of the sample is its aspect index, which is naturally available because we know which rationale is generated for which aspect. This

is where the *self-supervision* comes from. The self-supervised contrastive loss is calculated as in Equation 1 with the $K \cdot N$ samples in K classes and each class has N samples.

3.4 Three-Stage Training

Training the model directly with cross entropy or self-supervised contrastive loss suffers from the interlocking problem. We propose a three-stage training framework that alleviates the interlocking problem. As illustrated in Figure 2, the model is trained with the cross entropy and self-supervised contrastive loss together in the first stage. In the second stage, the generator is re-initialized, and is trained with the self-supervised contrastive loss L^{self} with the **predictor frozen** and only the generator can be updated. In the third stage, the model is trained with the cross entropy and the self-supervised contrastive loss again, with the **generator frozen** and only the predictor can be updated.

In the second stage, the objective to be optimized is no longer dominated by the same loss as in the first stage (cross entropy), and only the generator can be updated while the predictor is frozen, therefore they cannot be **inter-locked**. The generator learns to select rationales that are semantically far apart from each other with the contrastive loss. Intuitively, the shared encoder of the predictor is trained to learn the representation of rationales in the first stage, which are fully exploited in the second stage using a different loss. Similarly, the objective to be optimized in the third stage is not dominated by the contrastive loss as in the second stage, and only the predictor can be updated, therefore the generator and predictor cannot be **inter-locked**.

One may wonder why the cross entropy dominates in the first stage instead of the contrastive loss. In our situation, we have empirically observed that the cross entropy can be optimized to around the same value with or without the contrastive loss together, but the contrastive loss can be optimized to a significantly lower value alone without the cross entropy. This suggests that the cross entropy dominates the contrastive loss in our settings.

It is worth mentioning that it is essential to optimize different objectives in three stages. In the second stage, for example, if we optimize the same loss, it is still interlocked (or locked) because the objective to be optimized keeps the concavity with regard to the generator. Even worse, it gets less likely to jump out of the suboptimal than before

		Avg. Len.	Acc.	Avg. F1	App. F1	Aro. F1	Pal. F1	Tas. F1	Ove. F1
<i>Long</i>	Vanilla	35.5 / 25.1	91.7	42.4	57.7	37.0	26.7	29.2	61.4
	Contra	35.5 / 25.1	-	45.0	62.0	43.1	20.7	38.7	60.7
	3Stage	35.5 / 25.1	91.7	45.6	63.5	41.6	26.3	26.0	70.9
<i>Short</i>	Vanilla	23.7 / 18.8	90.9	34.4	37.0	31.4	21.3	33.2	26.9
	Contra	25.4 / 20.6	-	36.1	58.9	41.4	20.8	35.1	24.3
	3Stage	23.9 / 19.9	90.6	46.5	59.8	48.1	28.7	27.5	68.3

Table 1: Main results of different training methods on the two modes *Long* and *Short*. The average length is reported on both test set and annotation set. F1 is reported for all aspects described in section 4.1. **Bold** marks best.

because the predictor is frozen, which means the concave loss landscape for the generator becomes fixed and the generator can get stuck more easily.

4 Experiments

4.1 Datasets

We train and evaluate the classification and rationalization performance on the multi-aspect beer reviews dataset (McAuley et al., 2012). Each review describes five aspects related to beer: *appearance*, *aroma*, *palate*, *taste* and *overall*. For each aspect, a rating $\in \{0.2, 0.3, \dots, 0.9, 1.0\}$ is given (but our model uses only the overall rating). Following prior works (Bao et al., 2018), we binarized the ratings by considering ratings ≤ 0.4 as negative and ratings ≥ 0.6 as positive. The number of positive and negative samples are around the same. 60,000 balanced samples are sampled. There are 994 reviews with sentence-level aspect annotations.

4.2 Training Details

For all experiments, we used the 200-dimensional GloVe word embeddings (Pennington et al., 2014) trained on Wikipedia. We used the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.0001 and the batch size is 250. The temperature for contrastive loss is $\tau_c = 0.07$. All models are trained with a pre-defined number of epochs and the checkpoint with the minimum loss on validation is evaluated on the test set (for classification accuracy) and annotation set (for rationalization F1-score). Following prior works (Antognini and Faltings, 2021), we manually map the rationales to the aspect ordering that leads to the best F1-score.

4.3 Results

Table 1 show results on the beer reviews dataset obtained by evaluating the model described in Section 3.1 trained by three different methods: *Vanilla*, *Contra*, and *3Stage*. *Vanilla* is the baseline model

that adopts the single-stage training with the cross entropy loss, which is the most common loss in previous works (Lei et al., 2016; Antognini and Faltings, 2021). *Contra* also adopts the single-stage training but with the self-supervised contrastive loss solely. It is trained without labels, thus not able to do classification. *3Stage* is our proposed three-stage training method described in Section 3.4. All models are evaluated on two modes: *Long* and *Short*, which have different lengths of rationales on average. Specifically, in *Long* mode, the rationales are longer and all tokens are selected in one of the rationales, while in *Short* mode, the rationales are shorter. It is worth mentioning that *Short* models are trained in *Long* mode in the first stage so that a better representation can be learned with a wider horizon for the shared encoder.

Compared with *Vanilla* models, *Contra* can generate (slightly to moderately) better rationales in both *Short* and *Long* modes even without ground-truth information. This suggests that the self-supervised contrastive loss alone is a good objective for learning rationalization, thanks to its ability to promote semantic diversity among rationales. Compared with *Vanilla* models, *3Stage* achieves a moderately higher average F1-score in *Long* mode and significantly higher in *Short* mode, while reaching the same level of classification accuracy. Particularly, the F1-score is significantly higher in terms of the *overall* aspect in both modes. This suggests that *3Stage* can effectively jump out of suboptimal equilibrium and see the more informative text.

5 Conclusion

We proposed a multi-stage training method that avoids the interlocking problem. Its key ingredient is an additional contrastive loss that guides the learning of diverse rationales. In our multi-aspect scenario, experiments on the beer review dataset show that our method achieves significantly better rationalization.

References

- Diego Antognini and Boi Faltings. 2021. [Rationalization through concepts](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 761–775, Online. Association for Computational Linguistics.
- Diego Antognini, Claudiu Musat, and Boi Faltings. 2021. [Multi-dimensional explanation of target variables from documents](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12507–12515.
- 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, Brussels, Belgium. Association for Computational Linguistics.
- Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. [Interpretable neural predictions with differentiable binary variables](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.
- Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2019a. [A game theoretic approach to class-wise selective rationalization](#). *CoRR*, abs/1910.12853.
- Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2019b. *A Game Theoretic Approach to Class-Wise Selective Rationalization*. Curran Associates Inc., Red Hook, NY, USA.
- Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2020. [Invariant rationalization](#). *CoRR*, abs/2003.09772.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. [A simple framework for contrastive learning of visual representations](#). *CoRR*, abs/2002.05709.
- Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. [Very deep convolutional networks for text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, Valencia, Spain. Association for Computational Linguistics.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2019. [Momentum contrast for unsupervised visual representation learning](#). *CoRR*, abs/1911.05722.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. [Supervised contrastive learning](#). *CoRR*, abs/2004.11362.
- Yoon Kim. 2014. [Convolutional neural networks for sentence classification](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. [Rationalizing neural predictions](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.
- Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE.
- Jeffrey Pennington, Richard Socher, and Christopher 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, Doha, Qatar. Association for Computational Linguistics.
- Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. [Rethinking cooperative rationalization: Introspective extraction and complement control](#). 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 4094–4103, Hong Kong, China. Association for Computational Linguistics.
- Mo Yu, Yang Zhang, Shiyu Chang, and Tommi Jaakkola. 2021. Understanding interlocking dynamics of cooperative rationalization. *Advances in Neural Information Processing Systems*, 34.