SparCAssist: A Model Risk Assessment Assistant Based on Sparse Generated Counterfactuals

Zijian Zhang
L3S Research Center
Hannover, Lower Saxony, Germany

Vinay Setty
L3S Research Center
Germany
University of Stavanger
Stavanger, Norway

Avishek Anand
L3S Research Center
Germany
Delft University of Technology
Delft, Netherlands

ABSTRACT

We introduce SparCAssist, a general-purpose risk assessment tool for the machine learning models trained for language tasks. It evaluates models’ risk by inspecting their behavior on counterfactuals, namely out-of-distribution instances generated based on the given data instance. The counterfactuals are generated by replacing tokens in rational subsequences identified by ExPred, while the replacements are retrieved using HotFlip or Masked Language Model based algorithms. The main purpose of our system is to help the human annotators to assess the model’s risk on deployment. The counterfactual instances generated during the assessment are the by-product and can be used to train more robust NLP models in the future.

CCS CONCEPTS

- Information systems → Online analytical processing;
- Human-centered computing → User interface toolkits;
- Computing methodologies → Reasoning about belief and knowledge.

KEYWORDS

Interpretable Machine Learning, Counterfactual Interpretation, Data-annotation tools, Human-in-the-loop Machine Learning

ACM Reference Format:
Zijian Zhang, Vinay Setty, and Avishek Anand. 2022. SparCAssist: A Model Risk Assessment Assistant Based on Sparse Generated Counterfactuals. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’22), July 11–15, 2022, Madrid, Spain. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3531677

1 INTRODUCTION

In this paper we present a demonstration of SparCAssist: a tool for human-assisted debugging of machine learning (ML) models for language tasks. Debugging machine learning models in terms of estimating if a trained model is fit for deployment is considered more of an art than a science with a lot depending on the heuristics employed by the model developer [1]. Standard deployment checks often involve heavy reliance on performance metrics on validation and test sets to evaluate the generalization and robustness of systems. That means there is heavy focus on inspecting incorrectly predicted instances. We argue that ensuring good performance is not enough. Deep neural networks for language tasks are known to be particularly fragile on out-of-distribution instances even if they deliver impressive validation set accuracy [24].

Some of the known risks of deploying such over-parameterized models such as transformers is that they tend to memorize undesirable shortcuts in the training data that lead to reasonable validation performance but leads to failures in real-world scenarios [37]. Unlike common practice of ML metrics, we propose to provide a system that uses counterfactuals as examples to help the user, in this case the model developer, to understand the risks of the model under consideration. Specifically, given a trained ML model, our intention is to allow the user to identify instances that have potential risk during deployment. An instance is considered risky if its minor human-understandable modifications result in undesirable predictions.

Consider a document classification task of classifying review text of movies into positive and negative sentiments. Say a trained sentiment classifier classifies the input “the movie was awesome” as a positive sentiment. However, by slightly perturbing it to a counterfactual sentence “the movie was is awesome” results in a negative sentiment. This counterfactual sentence exposes two potential risks in the model, firstly, the model focuses on non-essential words towards making a prediction and secondly, the model is not robust under non-decisive perturbation. Other approaches that provide insights into the reasons behind predictions of complex machine learning models, a large variety of post-hoc explanation methods have been proposed [17, 22, 32]. However, all these explanation approaches operate on existing instances on the training or the test set. Therefore they cannot effectively assess the potential risk when the models are deployed in the field.

The aim of SparCAssist is to modify existing counterfactual generation approaches to help humans identify instances that might be particularly risky. We allow users enough control to choose input instances that are of interest and pieces of text out of them within that might be important for the task under consideration. In allowing a fine-grained control to the inputs choosing and feedback providing, the users are able to both get a sense of the functioning of the model as well as be educated about the potential risks it might possess. In SparCAssist, we work on the task of sentiment classification as an example but in principle SparCAssist can
extended to a large set of language tasks, for example question answering, fact checking, textual entailment etc.

The system then provides multiple “perturbations” of the input sentence that result in a change in the prediction. The perturbations should be small enough so that we can effectively assess the model’s local risk, specifically in the close vicinity surrounding a given instance. Therefore, we are only interested in sparse perturbations or we only allow a small number of perturbations to the original input. The users’ role is to assess whether the generated counterfactual is plausible, natural and if the model’s prediction is correct on counterfactuals. Plausible here means that the input text is believed to be written or conceived by a human. In terms of approach we start from a trained model and identify sentences that need to be scrutinized. We choose this subset of sentences using ExPred [35], a select-then-predict system that is interpretable by design. We then use adversarial approaches like HotFlip [5] and Masked Language Model approaches to generate counterfactuals. These generated counterfactuals are then shown to humans who in turn interact with the suggestions to establish its plausibility and validity. Finally, we compute the risk of the model under consideration from the trail of user interactions.

The primary goal is to assess the model’s deficiencies using the perturbations, and then our secondary goal is to generate a dataset of counterfactual explanations, which are grammatically plausible and meaningful from human perspective, could be the by-product of the system.

1.1 Related Work

We divide the related work into two parts – counterfactual generation, and automated approaches for model debugging. The counterfactual generation approaches can be categorized into two classes: automatic approaches and human-annotation approaches. Automatic approaches for generating counterfactuals are predominantly based on gradient-based optimization to generate adversarial examples like HotFlip [5], AutoPrompt [25], and Imitation Attacks [33]. Another set of approaches to automatically generate counterfactuals is to use prompt-based approaches like Polyjuice [34]. However, the main disadvantage of automatic approaches is that the generated counterfactuals are not natural, i.e., likely to be not plausible from human perspective. Prompt-based approaches are slightly more natural that gradient-based approaches, but they rely on heuristic prompts that have to be manually provided.

On the other hand, some approaches conduct user-study with crowd workers to obtain human-understandable counterfactual datasets [10, 18]. These approaches are understandably non-scalable and exhibit high variance due to subjectivity and the well-known Rashomon effect [3]. We use both gradient-based and prompt-based approaches for counterfactual generation but instead use humans-in-the-loop methods for quality assessment. More precisely, we take a hybrid approach for producing counterfactuals with utility of the counterfactual as the most desirable outcome.

The second piece of related work is regarding debugging ML models that have been used in text classification [21], search [27, 29, 31], and many language tasks in general [6, 9] by using explanations or interpretable machine learning approaches called explanation-based human debugging (EBHD). Lertvittayakumjorn and Toni [12] recently review EBHD approaches that exploit explanations to enable humans to give feedback and debug NLP models. Unlike adversarial algorithms [20, 33] the goals of these systems is to explicitly expose model’s deployment risk by ensuring that models are right for the right reason. In addition to academic efforts, model debugging tools like [28, 36], SageMaker 1 provided by Amazon are also available. Other human-in-the-loop systems include [8, 30] to update models and knowledge bases. However, they mostly focus on training process and do not provide a post-hoc model risk assessment.

2 SOLUTION APPROACH

The primary goal of our SparCAssist tool is to assist users to assess the deployment risk of the NLP classification model trained on specific datasets using counterfactual instances (counterfactuals). The input to SparCAssist is a model together with a dataset using which the risk assessment of the model is to be performed. The SparCAssist generates counterfactuals based on a selected data instance and tries to flip the model’s prediction for that instance. The tool assesses the risks of the model based on the user’s feedback on the plausibility and meaningfulness of the counterfactuals. The secondary output SparCAssist tool is plausible and meaningful counterfactual dataset for future training. In this section, we state our problem by first defining the counterfactual formally with its three essential properties and then formalize the model’s risk.

2.1 Interpretation Using Counterfactuals

Interpretation-by-Counterfactuals, also known as interpretation using minimal contrastives [24] is a model-interpretation technique that allows humans to understand the reason behind prediction by investigating models’ behavior under counterfactual inputs [14, 19]. According to [7, 24], in order to assess the risk of the models, the counterfactuals need to satisfy following rules:

1. The number of replaced tokens should be as few as possible.
2. The counterfactual instance should be plausible and meaningful [2].

Like in HotFlip [5] and Polyjuice [34], we mainly focus on the counterfactual generated by replacing the tokens. These two approaches, however, do not constrain the number of replaceable tokens while generating the counterfactuals. This will result in an overwhelming number of alternative counterfactuals for human annotators. In SparCAssist, we address this issue by firstly focusing on sentences containing rationale subsequences identified by the ExPred model [35]. We then only allow rationale tokens identified by ExPred in the selected sentence to be replaced, while the other tokens are kept fixed. Alternatively to the ExPred, users can manually select the tokens allowed to be replaced in the rationale sentences.

More formally, we use $x = [x_0, x_1, \ldots, x_l]$ to represent the input sequence (rationale sentence) to our tool and the model to be assessed is represented as $M$, while its predicted label for a given instance $x$ is denoted as $\hat{y} = M(x \odot m)$, where $m \in \{0, 1\}$ is the rationale mask, $\odot$ stands for element-wise multiplication. We then define a counterfactual generation function $C_{x_i \rightarrow y}$ which replaces

1https://docs.aws.amazon.com/sagemaker/latest/dg/train-debugger.html
the rationale token $x_i$ with a counterfactual token $v$ from the vocabulary $V$. A counterfactual with only one token replaced is called 1-counterfactual and represented as $c_i^v = C_{x_i\rightarrow v}(m) = (m \odot x)$. The condition $i \in \mathcal{P}(m)$ restricts the tokens to be replaced to the rational mask $m$. We repeat the above processes until the prediction on the counterfactual sentence changes.

### 2.2 Model Risk Assessment

We consider a model as “risky”, when its behavior on the counterfactuals does not match domain experts’ (users’) decision, despite its high accuracy on the given, or inside-distribution, instances, including instances from the train, dev-test, and test set [11]. Since the global risk is usually intractable [23], we consider its local counterpart instead, namely how wrong does the model predict on the counterfactuals sampled from the vicinity of an instance in the dataset.

As illustrated in Fig. 1, for a counterfactual $x^c$ sampled from the vicinity of an inside-distribution instance $x \in S(x)$, if the decision of the model $M(x^c)$ is dissonant with the decision of domain experts’, we call this an “risky behavior” and the area resulting in risky behavior “risky volume”. The local risk score of the model is the ratio between the volume of risky area and the volume of $S$. In more detail, we assume an “optimal” decision boundary $B^*$ as a hyperplane resulting in decisions aligning with domain experts’ prediction, while $B$ is the decision boundary of the model. The model’s local risk in $S$ is then the ratio between the volume of $V$ and $S$, where $V$ is the interior enclosed by $S$ and both decision boundaries as $V$, namely $\partial V = (B^* \cup B) \cap S$. We can estimate it by dividing the number of samples resulting in risky behaviors with the number of samples drawn universally from $S$ using the Monte-Carlo strategy. However, it is difficult to draw samples from such vicinity due to the discreetness of human language sentences. Therefore, instead of drawing counterfactual samples universally, we focus on such counterfactuals resulting in the model’s prediction flipping. Such examples locate right across the model’s decision boundary and will most likely expose the model’s risk.

For example, the sentiment classifier should predict $x = “I love this movie.”$ as positive and the counterfactuals $x_i^c = “I love hate this movie.”$ to negative. Because this is in accord with users’ decisions, the model’s risk does not increase. But if the model predicts $x_i^c = “I like this movie film.”$ to negative, the risk score increases, as users will not agree with this prediction. We ask the annotator to rate the prediction ($f \in 1-5, f$ for faithfulness) by asking to how much extent does the model predicts in concord with her expectation. We do not ask for a binary rating because the optimal decision boundaries are vague and subjective. Then for each annotator, we calculate her risk score around the instance $x$ by: $\text{risk} = \sum_{c \in X^c} (5-f(x^c))/|X^c|$, where $X^c$ represents the set of counterfactual samples provided by this annotator. The final risk of the model in $S$ is estimated by the weighted average of risk scores from all annotators, where the weight is the number of the counterfactuals generated in her/his session. It is worth mentioning that it requires further research to decide the exact number of counterfactuals needed to effectively evaluate the model’s risk.

### 3 SYSTEM DESIGN

As illustrated in Figure 2, we decompose the SparCAssist system into three components: Rationale Selection, Position Attribution and Word Replacement. The SparCAssist generates counterfactuals are as follows:

1. (1) It first randomly selects an instance $x$ from the given dataset which may or may not be labeled. For the demo, we use the “Movies” datasets for the sentiment analysis task [4].
2. (2) It then leverages the ExPred model [35] which is a two-phase select-then-predict model pre-trained on the movies dataset to select the rationale tokens.
3. (3) For all those rationale tokens, we attribute their positions and select candidate position(s) based on the attribution. In our SparCAssist, we are only interested in such counterfactuals changes the model’s prediction ($M(c) \neq \hat{y}$), therefore:
4. (4) It replaces the tokens on each selected position so that the estimated loss regarding the alternative label other than the original prediction is minimized.
5. (5) It then selects the final single counterfactual with only one token replaced based on their actual prediction loss.
6. (6) We repeat steps 3 to 5 until the prediction changes or the number of word replacements exceeds a pre-defined upper limit. Based on our empirical observation, the set the upper limit to five replacements.

We integrate two counterfactual generation approaches to SparCAssist, the Taylor-expansion based HotFlip and a Masked Language Model (MLM)-based algorithm used in Polyjuice. These two approaches are decomposed into two phases, namely position attribution and token selection as step 3 and 4 introduced above. We explain these two approaches below:

**HotFlip approach:** For this approach, we follow [5] to search for positions and words using the beam search. More formally, we attribute positions with the dot-product of corresponding token embedding with the partial derivative of the loss regarding the alternated prediction label with respect to the token embedding, i.e. $\frac{\partial}{\partial x_i}(\hat{y} \neq \hat{y}) \cdot e_i$, where $e_i$ represents the embedding of the $i$-th token. We select the top-$p$, $p \leq r$ of the position, where $r$ is the number of rationale tokens. Then in token selection, for
each select position we choose the token that maximizes the dot-product of the new token’s embedding with the partial derivation, i.e. 
\[ \text{replace} = \arg\max_{v \in V} \frac{\partial}{\partial x_{i,l}} (\hat{y}, \bar{\hat{y}} \neq \hat{y}) \text{e}^v. \]

**Masked Language Modelling-based approach:** For this method, we first construct the MLM model’s input with the prompt template as suggested in [13, 15, 26] as follows: “<masked sequence>” the sentiment of this review is “<alternative label>”. For example, “this movie is [mask]” the sentiment of this review is “negative”.

Unlike in the hotflip, where the positions are attributed with the product of gradient and the embedding, there is no such close-form position attribution in the MLM-based method. Therefore we trivially attribute all tokens with the same score For the token replacement, we mask each rationale word and apply the prompt template, then predict the masked word using a pre-trained blank-filling model based on RoBERTa [16] (with pre-trained weights from the Huggingface framework\(^2\)). For each word position, we keep the token with the highest score, then we select the final counterfactual which minimizes the loss with respect to the alternative label.

A screenshot of our system is illustrated in Fig. 3. The top part shows widgets to upload the model and the corresponding dataset. The lower part is split into two columns:

1. The left part (“1. Select a sentence”) presents a document selected from the dataset, where rationale tokens are highlighted with boldface font. By default, all sentences containing at least one rationale token are displayed. The users can expand the full review by clicking the “Expand all” button. Even after restricting the sentences using the rationale masks, the length of the whole review document is still large, users can therefore select a single sentence at a time to inspect its counterfactuals.

2. In the right hand side, (“2. How do you like the counterfactuals?”) the counterfactuals are displayed following the order of word replacement, with a replaced word highlighted. As introduced in the previous section, the users are asked to evaluate the plausibility and meaningfulness, together with the faithfulness of the prediction, if the user is satisfied with the current counterfactual generation. The users may also choose to use the default ExPred rationale masks or provide a custom mask by selecting the alternative rationale tokens according to their preference.

4 CONCLUSION AND FUTURE WORK

In this paper, we propose SparCAssist, an ML model debugging tool. SparCAssist allows users to construct word-replacement-based counterfactual instances of rationale masked sentences. These counterfactuals are helpful for better understanding of model’s behavior and identifying deployment risks of the underlying the model. The SparCAssist prunes the choices of counterfactuals by restricting the range of variable tokens to rationale tokens only. This minimizes the number of counterfactual choices presented to the users. Optionally, the users can also specify custom masks for the variable tokens. We implement HotFlip and Masked Language Model-based approaches to retrieve out-of-distribution (OOD) tokens and enable the users to evaluate generation results. It is also possible to deploy the SparCAssist system and gather counterfactual annotations with evaluations from large crowd-workers through web browsers. Another advantage of our tool is that it is flexible and generic. It allows debugging any ML model with a user chosen dataset. It easily supports replacing the rationale-masking, position attribution and OOD tokens selection components with other methods.
