Diagnostics-Guided Explanation Generation

Atanasova, Pepa Kostadinova; Simonsen, Jakob Grue; Lioma, Christina; Augenstein, Isabelle

Published in:
Proceedings of the International Joint Conference on Artificial Intelligence

DOI:
10.1609/aaai.v36i10.21287

Publication date:
2022

Document license:
Unspecified

Citation for published version (APA):
Abstract

Explanations shed light on a machine learning model’s rationales and can aid in identifying deficiencies in its reasoning process. Explanation generation models are typically trained in a supervised way given human explanations. When such annotations are not available, explanations are often selected as those portions of the input that maximise a downstream task’s performance, which corresponds to optimising an explanation’s Faithfulness to a given model. Faithfulness is one of several so-called diagnostic properties, which prior work has identified as useful for gauging the quality of an explanation without requiring annotations. Other diagnostic properties are Data Consistency, which measures how similar explanations are for similar input instances, and Confidence Indication, which shows whether the explanation reflects the confidence of the model. In this work, we show how to directly optimise for these diagnostic properties when training a model to generate sentence-level explanations, which markedly improves explanation quality, agreement with human rationales, and downstream task performance on three complex reasoning tasks.

1 Introduction

Explanations are an important complement to the predictions of a ML model. They unveil the decisions of a model that lead to a particular prediction, which increases user trust in the automated system and can help find its vulnerabilities. Moreover, “The right . . . to obtain an explanation of the decision reached” is enshrined in the European law (Regulation 2016).

In NLP, research on explanation generation has spurred the release of datasets (Zaidan, Eisner, and Piatko 2008; Thorne et al. 2018; Khashabi et al. 2018) containing human rationales for the correct predictions of downstream tasks in the form of word- or sentence-level selections of the input text. Such datasets are particularly beneficial for knowledge-intensive tasks (Petroni et al. 2020) with long sentence-level explanations, e.g., question answering and fact-checking, where identifying the required information is an important prerequisite for a correct prediction. They can be used to supervise and evaluate whether a model employs the correct

Figure 1: Example instance from MultiRC with predicted target and explanation (Step 1), where sentences with confidence $\geq 0.5$ are selected as explanations (S17, S18, S20). Steps 2-4 illustrate the use of Faithfulness, Data Consistency, and Confidence Indication diagnostic properties as additional learning signals. ‘[MASK](2)’ is used in Step 2 for sentences (in red) that are not explanations, and ‘[MASK](4)’–for random words in Step 4.

rationales for its predictions (DeYoung et al. 2020a; Thorne et al. 2018; Augenstein 2021). The goal of this paper is to improve the sentence-level explanations generated for such complex reasoning tasks.

When human explanation annotations are not present, a common approach (Lei, Barzilay, and Jaakkola 2016; Yu et al. 2019) is to train models that select regions from the input maximising proximity to original task performance which corresponds to the Faithfulness property. Atanasova et al. (2020a) propose Faithfulness and other diagnostic properties to evaluate different characteristics of explanations. These include Data Consistency, which measures the similarity of the explanations between similar instances, and Confidence Indication, which evaluates whether the expla-

*Contact Author

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
nation reflects the model’s confidence, among others (see Figure 1 for an example).

**Contributions**

We present the first method to learn the aforementioned diagnostic properties in an unsupervised way, directly optimising for them to improve the quality of generated explanations. We implement a joint task prediction and explanation generation model, which selects rationales at sentence level. Each property can then be included as an additional training objective in the joint model. With experiments on three complex reasoning tasks, we find that apart from improving the properties we optimised for, diagnostic-guided training also leads to explanations with higher agreement with human rationales, and improved downstream task performance. Moreover, we find that jointly optimising for diagnostic properties leads to reduced claim/question-only bias (Schuster et al. 2019) for the target prediction, and means that the model relies more extensively on the provided evidence. Importantly, we also find that optimising for diagnostic properties of explanations without supervision for explanation generation does not lead to good human agreement. This indicates the need for human rationales to train models that make the right predictions for the right reasons.

2 Related Work

**Supervised Explanations.** In an effort to guide ML models to perform human-like reasoning and avoid learning spurious patterns (Zhang, Marshall, and Wallace 2016; Ghaeini et al. 2019), multiple datasets with explanation annotations at the word and sentence level have been proposed (Wiegreffe and Marasovic 2021). These annotations are also used for supervised explanation generation, e.g., in pipeline models, where the generation task is followed by predicting the target task from the selected rationales only (DeYoung et al. 2020b; Lehman et al. 2019). As Wiegreffe, Marasović, and Smith (2020); Kumar and Talukdar (2020); Jacovi and Goldberg (2021) point out, pipeline models produce explanations without task-specific knowledge and without knowing the label to explain. However, for completion, we include the baseline pipeline from ERASER’s benchmark DeYoung et al. (2020b) as a reference model for our experiments.

Explanation generation can also be trained jointly with the target task (Atanasova et al. 2020b; Li et al. 2018), which has been shown to improve the performance for both tasks. Furthermore, Wiegreffe, Marasović, and Smith (2020) suggest that self-rationa-lising models, such as multi-task models, provide more label-informed rationales than pipeline models. Such multi-task models can additionally learn a joint probability of the explanation and the target task prediction on the input. This can be decomposed into first extracting evidence, then predicting the class based on it (Zhao et al. 2020; Zhou et al. 2019), or vice versa (Pruthi et al. 2020a). In this work, we also employ joint conditional training. It additionally provides a good testbed for our experiments with ablutions of supervised and diagnostic property objectives, which is not possible with a pipeline approach.

Most multi-task models encode each sentence separately, then combine their representations, e.g., with Graph Attention Layers (Zhao et al. 2020; Zhou et al. 2019). Glockner, Habernal, and Gurevych (2020) predict the target label from each separate sentence encoding and use the most confident sentence prediction as explanation, which also allows for unsupervised explanation generation. We consider Glockner, Habernal, and Gurevych (2020) as a reference model, as it is the only other work that reports results on generating explanations at the sentence level for three complex reasoning datasets from the ERASER benchmark (DeYoung et al. 2020a). It also outperforms the baseline pipeline model we include from ERASER. Unlike related multi-task models, we encode the whole input jointly so that the resulting sentence representations are sensitive to the wider document context. The latter proves to be especially beneficial for explanations consisting of multiple sentences. Furthermore, while the model of Glockner, Habernal, and Gurevych (2020) is limited to a fixed and small number (up to two) of sentences per explanation, our model can predict a variable number of sentences depending on the separate instances’ rationales.

**Unsupervised Explanations.** When human explanation annotations are not provided, model’s rationales can be explained with post-hoc methods based on gradients (Sundararajan, Taly, and Yan 2017), simplifications (Ribeiro et al. 2016), or teacher-student setups (Pruthi et al. 2020b). Another approach is to select input tokens that preserve a model’s original prediction (Lei et al. 2018; Yu et al. 2019; Bastings, Aziz, and Titov 2019; Paranjape et al. 2020), which corresponds to the Faithfulness property of an explanation. However, as such explanations are not supervised by human rationales, they do not have high overlap with human annotations (DeYoung et al. 2020a). Rather, they explain what a model has learned, which does not always correspond to correct rationales and can contain spurious patterns (Wang and Culotta 2020).

3 Method

We propose a novel Transformer (Vaswani et al. 2017) based model to jointly optimise sentence-level explanation generation and downstream task performance. The joint training provides a suitable testbed for our experiments with supervised and diagnostic property objectives for a single model. The joint training optimises two training objectives for the two tasks at the same time. By leveraging information from each task, the model is guided to predict the target task based on correct rationales and to generate explanations based on the model’s information needs for target prediction. This provides additional useful information for training each of the tasks. Conducting joint training for these two tasks was shown to improve the performance for each of them (Zhao et al. 2020; Atanasova et al. 2020b).

The core novelty is that the model is trained to improve the quality of its explanations by using diagnostic properties of explanations as additional training signals (see Figure 1). We select the properties Faithfulness, Data Consistency, and
Confidence Indication, as they can be effectively formulated as training objectives. Faithfulness is also employed in explainability benchmarks (DeYoung et al. 2020a) and in related work for unsupervised token-level explanation generation (Lei, Barzilay, and Jaakkola 2016; Lei et al. 2018), whereas we consider it at sentence level. Further, multiple studies (Yeh et al. 2019; Alvarez-Melis and Jaakkola 2018) find that explainability techniques are not robust to insignificant and/or adversarial input perturbations, which we address with the Data Consistency property. We do not consider Human Agreement and Rationale Consistency, proposed in Atanasova et al. (2020a). The supervised explanation generation training employs human rationale annotations and thus addresses Human Agreement. Rationale Consistency requires the training of a second model, which is resource-expensive. Another property to investigate in future work is whether a model’s prediction can be simulated by another model trained only on the explanations (Hase et al. 2020; Treviso and Martins 2020; Pruhi et al. 2020c), which also requires training an additional model. We now describe each component in detail.

3.1 Joint Modelling

Let \( D = \{(x_i, y_i, e_i) | i \in [1, \#(D)]\} \) be a classification dataset. The textual input \( x_i = (q_i, a_i^{opt}, s_i) \) consists of a question or a claim, an optional answer, and several sentences (usually above 10) \( s_i = \{s_{ij} | j \in [1, \#(s_i)]\} \) used for predicting a classification label \( y_i \in [1, N] \). Additionally, \( D \) contains human rationale annotations selected from the sentences \( s_i \) as a binary vector \( e_i = \{e_{ij} \in \{0, 1\} | j \in [1, \#(s_i)]\} \), which defines a binary classification task for explanation extraction.

First, the joint model takes \( x_i \) as input and encodes it using a Transformer model, resulting in contextual token representations \( h^L = encode(x_i) \) from the final Transformer layer \( L \). From \( h^L \), we select the representations of the CLS token that precedes the question—as it is commonly used for downstream task prediction in Transformer architectures—and the CLS token representations preceding each sentence in \( s_i \), which we use for selecting sentences as an explanation. The selected representations are then transformed with two separate linear layers \( h^C \) for predicting the target, and \( h^E \) for generating the explanations, which have the same hidden size as the size of the contextual representations in \( h^L \).

Given representations from \( h^E \), a N-dimensional linear layer predicts the importance \( p^E \in \mathbb{R}^{#(s_i)} \) of the evidence sentences for the prediction of each class. As a final sentence importance score, we only take the score for the predicted class \( p^{E[c]} \) and add a sigmoid layer on top for predicting the binary explanation selection task. Given representations from \( h^C \), a N-dimensional linear layer with a soft-max layer on top predicts the target label \( p^C \in \mathbb{R} \). The model then predicts the joint conditional likelihood \( L \) of the target task and the generated explanation given the input (Eq. 1). This is factorised further into first extracting the explanations conditioned on the input and then predicting the target label (Eq. 2) based on the extracted explanations (assuming \( y_i \perp x_i | e_i \)).

\[
L = \prod_{i=1}^{\#(D)} p(y_i, e_i | x_i) \tag{1}
\]

\[
L = \prod_{i=1}^{\#(D)} p(e_i | x_i) p(y_i | e_i) \tag{2}
\]

We condition the label prediction on the explanation prediction by multiplying \( p^C \) and \( p^E \), resulting in the final prediction \( p^C \in \mathbb{R} \). The model is trained to optimise jointly the task target cross-entropy loss function (\( \mathcal{L}_C \)) and the explanation generation cross-entropy loss function (\( \mathcal{L}_E \)):

\[
\mathcal{L} = \mathcal{L}_C(p^C, y) + \mathcal{L}_E(p^{E[c]}, e) \tag{3}
\]

All loss terms of the diagnostic explainability properties described below are added to \( \mathcal{L} \) without additional hyperparameter weights for the separate loss terms.

3.2 Faithfulness (F)

The Faithfulness property guides explanation generation to select sentences preserving the original prediction, (Step 2, Fig. 1). In more detail, we take sentence explanation scores \( p^{E[c]} \in [0, 1] \) and sample from a Bernoulli distribution the sentences which should be preserved in the input: \( c^E \sim \text{Bern}(p^{E[c]}) \). Further, we make two predictions—one, where only the selected sentences are used as an input for the model, thus producing a new target label prediction \( l^S \), and one where we use only unselected sentences, producing the new target label prediction \( l^{Co} \). The assumption is that a high number \#(s = S) \#(l = S) matching the original \( l^C \) indicates the sufficiency (S) of the selected explanation. On the contrary, a low number \#(l = Co) \#(l = Co) matching the original \( l^C \) indicate the selected explanation is complete (Co) and no sentences indicating the correct label are missed. We then use the REINFORCE (Williams 1992) algorithm to maximise the reward:

\[
\mathcal{R}^F = \#(l^{C = S}) - \#(l^{C = Co}) - \|c^E - \lambda\| \tag{4}
\]

The last term is an additional sparsity penalty for selecting more/less than \( \lambda \%) of the input sentences as an explanation, \( \lambda \) is a hyper-parameter.

3.3 Data Consistency (DC)

Data Consistency measures how similar the explanations for similar instances are. Including it as an additional training objective can serve as regularisation for the model to be consistent in the generated explanations. To do so, we mask \( K \) random words in the input, where \( K \) is a hyper-parameter depending on the dataset. We use the masked text (M) as an input for the joint model, which predicts new sentence scores \( p^{EM} \). We then construct an \( \mathcal{L} \) loss term for the property to minimise for the absolute difference between \( p^E \) and \( p^{EM} \):

\[
\mathcal{L}_{DC} = |p^E - p^{EM}| \tag{5}
\]

We use \( \mathcal{L} \) instead of \( \mathcal{L} \) loss as we do not want to penalise for potentially masking important words, which would result in entirely different outlier predictions.
3.4 Confidence Indication (CI)

The CI property measures whether generated explanations reflect the confidence of the model’s predictions (Step 3, Fig. 1). We consider this a useful training objective to re-calibrate and align the prediction confidence values of both tasks. To learn explanations that indicate prediction confidence, we aggregate the sentence importance scores, taking their maximum, minimum, mean, and standard deviation. We transform the four statistics with a linear layer that predicts the confidence \( \hat{p}^C \) of the original prediction. We train the model to minimise \( \mathcal{L}_1 \) loss between \( p^C \) and \( \hat{p}^C \):

\[
\mathcal{L}_{CI} = |p^C - \hat{p}^C|
\]

We choose \( \mathcal{L}_1 \) as opposed to \( \mathcal{L}_2 \) loss as we do not want to penalise possible outliers due to sentences having high confidence for the opposite class.

4 Experiments

4.1 Datasets

We perform experiments on three datasets from the ERASER benchmark (DeYoung et al. 2020a) (FEVER, MultiRC, Movies), all of which require complex reasoning and have sentence-level rationales. For FEVER (Thorner et al. 2018), given a claim and an evidence document, a model has to predict the veracity of a claim \( \in \{\text{support, refute}\} \). The evidence for predicting the veracity has to be extracted as explanation. For MultiRC (Khashabi et al. 2018), given a question, an answer option, and a document, a model has to predict if the answer is correct. For Movies (Zaidan, Eisner, and Piatko 2008), the sentiment \( \in \{\text{positive, negative}\} \) of a long movie review has to be predicted. For Movies, as in Glockner, Habernal, and Gurevych (2020), we mark each sentence containing annotated explanation at token level as an explanation. Note that, in knowledge-intensive tasks such as fact checking and question answering also explored here, human rationales point to regions in the text containing the information needed for prediction. Identifying the required information becomes an important preliminary for the correct prediction rather than a plausibility indicator (Jacovi and Goldberg 2020), and is evaluated as well (e.g., FEVER score, Joint Accuracy).

4.2 Metrics

We evaluate the effect of using diagnostic properties as additional training objectives for explanation generation. We first measure their effect on selecting human-like explanations by evaluating precision, recall, and macro F1-score against human explanation annotations provided in each dataset (§5.1). Second, we compute how generating improved explanations affects the target task performance by computing accuracy and macro F1-score for the target task labels (§5.2). Additionally, as identifying the required information in knowledge-intensive datasets, such as FEVER and MultiRC, is an important preliminary for a correct prediction, and following Thorne et al. (2018); Glockner, Habernal, and Gurevych (2020), we evaluate the joint target and explanation performance by considering a prediction as correct only when the whole explanation is retrieved (Acc. Full). In case of multiple possible explanations \( e_i \) for one instance (ERASER provides comprehensive explanation annotations for the test sets), selecting one of them counts as a correct prediction. Finally, as diagnostic property training objectives target particular properties, we measure the improvements for each property (§5.3).

4.3 Experimental Setting

Our core goal is to measure the relative improvement of the explanations generated by the underlying model with (as opposed to without) diagnostic properties. We conduct experiments for the supervised model (Sup.), including separately Faithfulness (F), Data Consistency (DC), and Confidence Indication (CI), as well as all three (All) as additional training signals (§3).

Nevertheless, we include results from two other architectures generating sentence-level explanations that serve as a reference for explanation generation performance on the employed datasets. Particularly, we include the best supervised sentence explanation generation results reported in Glockner, Habernal, and Gurevych (2020), and the baseline pipeline model from ERASER, which extracts one sentence as explanation and uses it for target prediction (see §2 for a detailed comparison). We also include an additional baseline comparison for the target prediction task. The BERT Blackbox model predicts the target task from the whole document as an input without being supervised by human rationales. The results are as reported by Glockner, Habernal, and Gurevych (2020). In our experiments, we use BERT (Devlin et al. 2019) base-uncased as our base architecture, following Glockner, Habernal, and Gurevych (2020).

5 Results

5.1 Explanation Generation Results

In Table 1, we see that our supervised model performs better than Glockner, Habernal, and Gurevych (2020); DeYoung et al. (2020b). For the MultiRC dataset, where the explanation consists of more than one sentence, our model brings an improvement of more than 30 F1 points over the reference models, confirming the importance of the contextual information, which performs better than encoding each explanation sentence separately.

When using the diagnostic properties as additional training objectives, we see further improvements in the generated explanations. The most significant improvement is achieved with the Data Consistency property for all datasets with up to 2.5 F1 points over the underlying supervised model. We assume that the Data Consistency objective can be considered as a regularisation for the model’s instabilities at the explanation level. The second highest improvement is achieved with the Faithfulness property, increasing F1 by up to 1 F1 point for Movies and MultiRC. We assume that the property does not result in improvements for FEVER as it has multiple possible explanation annotations for one instance, which can make the task of selecting one sentence as a complete explanation ambiguous. Confidence Indication results in improvements only on Movies. We conjecture that Confidence Indication is the least related to promoting similar-
Table 1: Target task prediction (F1-C, Accuracy-C) and explanation generation (Precision-E, Recall-E, F1-E) results (mean and standard deviation over three random seed runs). Last columns measures joint prediction of target accuracy and explanation generation. The property with the best relative improvement over the supervised model is in bold.
Correlations. We further study this in Sec. 6.1. That this might also be due to the model learning spurious explanations, it is still possible to predict the same label based on the Faithfulness property. The latter phenomenon can be explained by the low joint performance of both tasks. We also see that even when removing the selected explanations, it is still possible to predict the same label based on the Faithfulness objective.

In Table 2 we see that supervision from the Faithfulness leads to generating explanations that preserve the original label of the instance for all datasets. For FEVER, the label is even preserved in 12% of the instances more than with the supervised objective only. The least faithful explanations are those generated for MultiRC, which can be explained by the low joint performance of both tasks. We also see that even when removing the selected explanations, it is still possible to predict the same label based on the remaining evidence. Such cases are decreased when including the Faithfulness property. The latter phenomenon can be explained by the fact that FEVER and Movies’ instances contain several possible explanations. We conjecture that this might also be due to the model learning spurious correlations. We further study this in Sec. 6.1.

Faithfulness. In Table 2 we see that supervision from the Faithfulness property leads to generating explanations that preserve the original label of the instance for all datasets. For FEVER, the label is even preserved in 12% of the instances more than with the supervised objective only. The least faithful explanations are those generated for MultiRC, which can be explained by the low joint performance of both tasks. We also see that even when removing the selected explanations, it is still possible to predict the same label based on the remaining evidence. Such cases are decreased when including the Faithfulness property. The latter phenomenon can be explained by the fact that FEVER and Movies’ instances contain several possible explanations. We conjecture that this might also be due to the model learning spurious correlations. We further study this in Sec. 6.1.

Data Consistency. Using Data Consistency as an additional training objective aims to regularise the model to select similar explanations for similar instances. In Table 3, we find the variance of downstream task prediction confidence decreases for all datasets with up to 0.04 points. Furthermore, the variance of generated explanation probabilities for similar instances is decreased as well. The largest improvements are for MultiRC and Movies, where the property brings the highest performance improvement w.r.t. human annotations as well. We also find that the Movies dataset, which has the longest inputs, has the smallest variance in explanation predictions. This suggests that the variance in explanation prediction is more pronounced for shorter inputs as in FEVER and MultiRC, where the property brings more improvement w.r.t. human annotations. The variance could also depend on the dataset’s nature.

Confidence Indication. Table 4 shows the difference between the confidence of the predicted target label and the confidence of the explanation sentence with the highest importance. Including Confidence Indication as a training objective indeed decreases the distance between the confidence of the two tasks, making it easier to judge the confidence of the model only based on the generated explanation’s confidence. The confidence is most prominently improved for the Movies dataset, where it is also the dataset with the largest improvements for supervised explanation generation with Confidence Indication objective.

Unsupervised Rationale Generation

We explore how well explanations can be generated without supervision from human explanation annotations. Table 5 shows that the performance of the unsupervised rationales is limited with an up to 47 F1 point decrease for FEVER compared to the supervised model. We assume that as our model encodes the whole input together, this leads to a uniform importance of all sentences as they share information through their context. While joint encoding improves the target prediction for complex reasoning datasets especially with more than one explanation sentence, this also limits the unsupervised learning potential of our architecture. As the model is not supervised to select explanations close to human ones, improving the diagnostic properties has a limited effect in improving the results w.r.t. human annotations.

6 Discussion

6.1 Question/Claim Only Bias

Prior work has found that models can learn spurious correlations between the target task and portions of the input text,
The second example from FEVER illustrates the effect of including Faithfulness as an objective. Naturally, for instances classified correctly by the supervised model, their generated explanation is improved to reflect the rationale used to predict the target. However, when the prediction is incorrect, the effect of the Faithfulness property is limited.

Finally, we find Confidence Indication often re-calibrates the prediction probabilities of generated explanations and predicted target tasks, which does not change many target predictions. This explains its limited effect as an additional training objective. The re-calibration also influences downstream task prediction confidence, as in the last example from the Movies dataset. This is a side effect of optimising the property while training the target task, where both explanation and target prediction confidence can be changed to achieve better alignment.

### 7 Conclusion

In this paper, we study the use of diagnostic properties for improving the quality of generated explanations. We find that including them as additional training objectives improves downstream task performance and generated explanations w.r.t. human rationale annotations. Moreover, using only the diagnostic properties as training objectives does not lead to a good performance compared to only using human rationale annotations. The latter indicates the need for human rationale annotations for supervising a model to base its predictions on the correct rationales. In future, we plan to experiment with application tasks with longer inputs, where current architectures have to be adjusted to make it computationally possible to encode longer inputs.
Acknowledgments

The research documented in this paper has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199. Isabelle Augenstein’s research is further partially funded by a DFF Sapere Aude research leader grant.

References


