Diagnostics-Guided Explanation Generation

Pepa Atanasova*, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein

Department of Computer Science, University of Copenhagen, Denmark
{pepa, simonsen, c.lioma, augenstein}@di.ku.dk

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
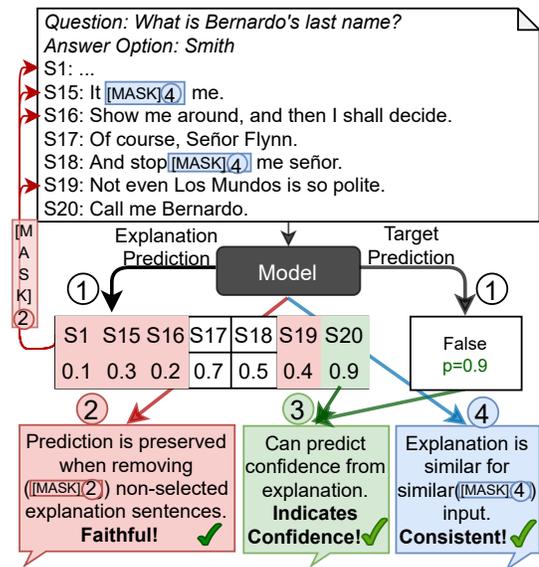
Figure 1: Example instance from MultiRC with predicted target and explanation (Step 1), where sentences with confidence ≥ 0.5 are selected as explanations (S17, S18, S20). Steps 2-4 illustrate the use of Faithfulness, Data Consistency, and Confidence Indication diagnostic properties as additional learning 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-

nation reflects the model's confidence, among others (see Figure 1 for an example).

Contributions[1] 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-rationalising 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*

---

[1]We make an extended version of the manuscript and code available on https://github.com/copenlu/diagnostic-guided-explanations .

ablations 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; Pruthi 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) $s_i = \{s_{i,j} | j \in [1, \#(s_i)]\}$ 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 $e_i = \{e_{i,j} = \{0,1\} | j \in [1, \#(s_i)]\}$, 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

$y_i \perp \mathbf{x}_i \mid \mathbf{e}_i)$.

$$L = \prod_{i=1}^{\#(D)} p(y_i, \mathbf{e}_i \mid \mathbf{x}_i) \tag{1}$$

$$L = \prod_{i=1}^{\#(D)} p(\mathbf{e}_i \mid \mathbf{x}_i) p(y_i \mid \mathbf{e}_i) \tag{2}$$

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 target task 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: $c^E \sim Bern(p^{E[c]})$. 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 $\#(l^{C=S})$ of predictions l^S matching the original l^C indicate the sufficiency (S) of the selected explanation. On the contrary, a low number $\#(l^{C=Co})$ of predictions 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, λ 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 $\mathcal{L}1$ 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 $\mathcal{L}1$ instead of $\mathcal{L}2$ 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 $\hat{p}^C$ and p^C:

$$\mathcal{L}_{CI} = |p^C - \hat{p}^C| \qquad (6)$$

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 (Thorne et al. 2018), given a claim and an evidence document, a model has to predict the veracity of a claim$\in$\{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$\in$\{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-

Dataset	Method	F1-C	Acc-C	P-E	R-E	F1-E	Acc-Joint
FEVER	Blackbox (Glockner, Habernal, and Gurevych 2020)	90.2 ±0.4	90.2 ±0.4				
	Pipeline (DeYoung et al. 2020a)	87.7	87.8	88.3	87.7	88.0	78.1
	Supervised (Glockner, Habernal, and Gurevych 2020)	90.7 ±0.7	90.7 ±0.7	92.3 ±0.1	91.6 ±0.1	91.9 ±0.1	83.9 ±0.1
	Supervised	89.3 ±0.4	89.4 ±0.3	94.0 ±0.1	93.8 ±0.1	93.9 ±0.1	80.1 ±0.4
	Supervised+Data Consistency	**89.7** ±0.5	**89.7** ±0.5	**94.4** ±0.0	**94.2** ±0.0	**94.4** ±0.0	80.8 ±0.5
	Supervised+Faithfulness	89.5 ±0.4	89.6 ±0.4	92.8 ±0.2	93.7 ±0.2	93.3 ±0.2	75.4 ±0.3
	Supervised+Confidence Indication	87.9 ±1.0	87.9 ±1.0	93.9 ±0.1	93.7 ±0.1	93.8 ±0.1	78.5 ±0.9
	Supervised+All	89.6 ±0.1	89.6 ±0.1	94.4 ±0.1	94.2 ±0.1	94.3 ±0.1	**80.9** ±0.1
MultiRC	Blackbox (Glockner, Habernal, and Gurevych 2020)	67.3 ±1.3	67.7 ±1.6				
	Pipeline (DeYoung et al. 2020a)	63.3	65.0	66.7	30.2	41.6	0.0
	Supervised (Glockner, Habernal, and Gurevych 2020)	65.5 ±3.6	67.7 ±1.5	65.8 ±0.2	42.3 ±3.9	51.4 ±2.8	7.1 ±2.6
	Supervised	71.0 ±0.3	71.4 ±0.3	78.0 ±0.1	78.6 ±0.5	78.3 ±0.1	16.2 ±0.4
	Supervised+Data Consistency	**71.7** ±0.6	**72.2** ±0.7	**79.9** ±0.4	79.0 ±0.8	79.4 ±0.5	**19.3** ±0.4
	Supervised+Faithfulness	71.0 ±0.4	71.3 ±0.4	78.2 ±0.1	79.1 ±0.2	78.6 ±0.1	16.1 ±0.5
	Supervised+Confidence Indication	70.6 ±0.7	71.1 ±0.6	77.9 ±0.8	78.3 ±0.5	78.1 ±0.5	16.5 ±1.0
	Supervised+All	70.5 ±1.6	71.2 ±1.3	79.7 ±1.1	**79.4** ±0.5	**79.6** ±0.7	18.8 ±1.6
Movies	Blackbox (Glockner, Habernal, and Gurevych 2020)	90.1 ±0.3	90.1 ±0.3				
	Pipeline (DeYoung et al. 2020a)	86.0	86.0	87.9	60.5	71.7	40.7
	Supervised (Glockner, Habernal, and Gurevych 2020)	85.6 ±3.6	85.8 ±3.5	86.9 ±2.5	62.4 ±0.1	72.6 ±0.9	43.9 ±0.6
	Supervised	87.4 ±0.4	87.4 ±0.4	79.6 ±0.6	68.9 ±0.5	73.8 ±0.5	59.4 ±0.6
	Supervised+Data Consistency	**90.0** ±0.7	**90.0** ±0.7	79.5 ±0.1	69.2 ±0.7	74.0 ±0.8	60.8 ±1.7
	Supervised+Faithfulness	89.1 ±0.6	89.1 ±0.6	**80.9** ±0.9	**69.9** ±1.3	**74.9** ±1.1	**62.6** ±1.6
	Supervised+Confidence Indication	89.9 ±0.7	89.9 ±0.7	79.7 ±1.4	69.5 ±0.7	74.3 ±1.0	60.1 ±2.6
	Supervised+All	89.9 ±0.7	89.9 ±0.7	80.0 ±1.0	69.5 ±1.0	74.4 ±1.0	60.3 ±2.2

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.

ity to human rationales in the generated explanations. Moreover, the re-calibration of the prediction confidence for both tasks possibly leads to fewer prediction changes, explaining the low scores w.r.t. human annotations. We look into how Confidence Indication affects the selected annotations in §5.3, and §6. Finally, combining all diagnostic property objectives, results in a performance close to the best performing property for each dataset.

5.2 Target Prediction Results

In Table 1, the Supervised model, without additional property objectives, consistently improves target task performance by up to 4 points in F1, compared to the two reference models that also generate explanations, except for FEVER, where the models already achieve high results. This can be due to the model encoding all explanation sentences at once, which allows for a more informed prediction of the correct target class. Our model trained jointly with the target task and explanation prediction objective also has similar performance to the BERT Blackbox model and even outperforms it by 4.4 F1 points for the MultiRC dataset. Apart from achieving high target prediction performance (F1-C) on the target task, our supervised model also learns which parts of the input are most important for the prediction, which is an important prerequisite for knowledge-intensive tasks.

We see further improvements in downstream task performance when using the diagnostic properties as additional training objectives. Improvements of the generated expla-

nations usually lead to improved target prediction as they are conditioned on the extracted evidence. Here, we again see that Data Consistency steadily improves the target task's performance with up to 2.5 F1 points. We also see improvements in F1 with Faithfulness for FEVER and MultiRC. Finally, we find that improvements in Confidence Indication lead to an improvement for target prediction of 2.5 F1 points for Movies. Combining all objectives, results in performance close the performance of the other properties.

We also show joint prediction results for target task and evidence. For MultiRC and Movies, the improvements of our supervised model over Glockner, Habernal, and Gurevych (2020) are very considerable with up to 9 accuracy points; using diagnostic properties increases results further up to 4 points in accuracy. Apart from improving the properties of the generated explanations, this could be due to the architecture conditioning the prediction on the explanation. The only dataset we do not see improvements for is FEVER, where again the performance is already high, and the target prediction of our model performs worse than Glockner, Habernal, and Gurevych (2020).

5.3 Explanations Property Results

So far, we concentrate on the relative performance improvements compared to human annotations. However, the diagnostic properties' additional training objectives are directed at generating explanations that exhibit these properties to a larger degree. Here, we demonstrate the improvements over

Dataset	Method	Suff. ↑	Compl. ↓
FEVER	Supervised	85.1	85.1
	Supervised+F	97.4	83.6
MultiRC	Supervised	81.7	69.2
	Supervised+F	82.3	67.0
Movies	Supervised	94.8	92.2
	Supervised+F	96.6	91.3

Table 2: Sufficiency and Completeness as proportions of the instances that preserve their prediction when evaluated on only the selected (Suff.) or the unselected (Compl.) explanation sentences, accordingly, for training with and without the Faithfulness objective.

Dataset	Method	Pred.	Expl.
FEVER	Sup.	0.03 (9.9e-8)	3.68 (1.80)
	Sup.+DC	0.02 (9.1e-8)	2.56 (0.97)
MultiRC	Sup.	0.09 (5.6e-8)	7.83(2.87)
	Sup.+DC	0.05 (4.9e-8)	3.01(0.89)
Movies	Sup.	0.04 (7.1e-8)	2.34 (1.38)
	Sup.+DC	0.01 (6.2e-8)	1.72 (0.90)

Table 3: Mean and standard deviation (in brackets) of the difference between target (Pred.) and explanation (Expl.) prediction confidence for similar (masked) instances.

the explanation properties themselves for unseen instances in the test splits. Note that this is a control experiment as we expect the properties we optimise for to be improved.

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

Method	FEVER	MultiRC	Movies
Sup.	0.10 (0.17)	0.05 (0.10)	0.12 (0.09)
Sup.+CI	0.05 (0.09)	0.04 (0.09)	0.05 (0.10)

Table 4: Mean and standard deviation (in brackets) difference between the model's confidence and the confidence of the generated explanations.

Method	FEVER	MultiRC	Movies
Sup.	93.9 $\pm_{0.1}$	78.3 $\pm_{0.1}$	73.8 $\pm_{0.5}$
UnS.	56.1 $\pm_{0.4}$	34.8 $\pm_{7.6}$	50.0 $\pm_{1.8}$
UnS.+DC	46.9 $\pm_{0.4}$	38.1 $\pm_{3.2}$	63.8 $\pm_{1.2}$
UnS.+F	51.6 $\pm_{0.3}$	24.4 $\pm_{5.2}$	64.6 $\pm_{0.4}$
UnS.+CI	57.5 $\pm_{0.4}$	25.4 $\pm_{3.4}$	60.0 $\pm_{1.6}$
UnS.+All	57.3 $\pm_{0.2}$	37.4 $\pm_{6.4}$	63.6 $\pm_{0.3}$

Table 5: Performance on the explanation generation task without human annotation supervision (UnS.).

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.

5.4 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,

Dataset	Method	F1-C	Acc-C
FEVER	Random	26.1 ±4.3	37.1 ±5.6
	Sup.	75.6 ±0.3	75.7 ±0.3
	Sup.+DC	68.2 ±0.2	75.6 ±0.3
	Sup.+F	73.4 ±0.4	73.9 ±0.3
	Sup.+CI	73.2 ±0.4	73.7 ±0.4
	Sup.+All	73.5 ±0.2	73.8 ±0.4
	Sup. on whole input	89.3 ±0.4	89.4 ±0.3
MultiRC	Random	26.1 ±5.5	31.6 ±5.9
	Sup.	59.4 ±0.8	63.5 ±0.9
	Sup.+DC	54.5 ±0.9	61.3 ±1.2
	Sup.+F	57.8 ±0.8	61.4 ±0.6
	Sup.+CI	49.7 ±0.8	60.1 ±0.2
	Sup.+All	59.0 ±0.3	61.0 ±0.2
	Sup. on whole input	71.0 ±0.3	71.4 ±0.3

Table 6: Performance of the models for the downstream task when provided with the query-answer part only.

e.g., predicting solely based on the claim to be fact checked (Schuster et al. 2019), regardless of the provided evidence. In our experiments, the input for FEVER and MultiRC also contains two parts - a claim or a question-answer pair and evidence text, where the correct prediction of the target always depends on the evidence. Suppose the models do not consider the second part of the input when predicting the target task. In that case, efforts to improve the generated explanations will not affect the target task prediction as it does not rely on that part of the input.

Table 6 shows target task performance of models trained on the whole input, but using only the first part of the input at test time. We find that, given the limited input, the performance is still considerable compared to a random prediction. For FEVER, the performance drops only with 14 F1-score points to 75.6 F1-score. This could explain the small relative improvements for FEVER when including diagnostic properties as training objectives, where the prediction does not rely on the explanation to a large extent.

Another interesting finding is that including diagnostic properties as training objectives decreases models' performance when a supporting document is not provided. We assume this indicates the properties guide the model to rely more on information in the document than to learn spurious correlations between the question/claim and the target only. The Data Consistency and Confidence Indication property lead to the largest decrease in model's performance on the limited input. This points to two potent objectives for reducing spurious correlations.

6.2 Explanation Examples

Table 7 illustrates common effects of the diagnostic properties. We find Data Consistency commonly improves explanations by removing sentences unrelated to the target prediction, as in the first example from MultiRC. This is particularly useful for MultiRC, which has multiple gold explanation sentences. For FEVER and Movies, where one sentence is needed, the property brings smaller improvements w.r.t. human explanation annotations.

Question: What colors are definitely used in the picture Lucy drew?; **Answer:** Yellow and purple; **Label:** True
Predicted: Sup True, p=.98; **Sup+DC** True, p=.99
E-Sup: She draws a picture of her family. She makes sure to draw her mom named Martha wearing a purple dress, because that is her favorite. She draws many yellow feathers for her pet bird named Andy.
E-Sup+S: She makes sure to draw her mom named Martha wearing a purple dress, because that is her favorite. She draws many yellow feathers for her pet bird named Andy.

Claim: Zoey Deutch did not portray Rosemarie Hathaway in Vampire Academy.; **Label:** REFUTE
Predicted: Sup refute, p=.99; **Sup+F** refute, p=.99
E-Sup: Zoey Francis Thompson Deutch (born November 10, 1994) is an American actress.
E-Sup+F: She is known for portraying Rosemarie "Rose" Hathaway in Vampire Academy(2014), Beverly in the Richard Link later film Everybody Wants Some!!

E-Sup/E-Sup+CI: For me, they calibrated my creativity as a child; they are masterful, original works of art that mix moving stories with what were astonishing special effects at the time (and they still hold up pretty well).; **Label:** Positive
Predicted: Sup negative, p=.99 **Sup+CI** positive, p=.99

Table 7: Example explanation predictions changed by including the diagnostic properties as training objectives.

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

References

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Atanasova, P.; Simonsen, J. G.; Lioma, C.; and Augenstein, I. 2020a. A Diagnostic Study of Explainability Techniques for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3256–3274. Online: Association for Computational Linguistics.

Atanasova, P.; Simonsen, J. G.; Lioma, C.; and Augenstein, I. 2020b. Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7352–7364. Online: Association for Computational Linguistics.

Augenstein, I. 2021. *Towards Explainable Fact Checking*. Dr. Scient. thesis, University of Copenhagen, Faculty of Science.

Bastings, J.; Aziz, W.; and Titov, I. 2019. Interpretable Neural Predictions with Differentiable Binary Variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2963–2977. Florence, Italy: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020a. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458. Online: Association for Computational Linguistics.

DeYoung, J.; Lehman, E.; Nye, B.; Marshall, I.; and Wallace, B. C. 2020b. Evidence Inference 2.0: More Data, Better Models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 123–132. Online: Association for Computational Linguistics.

Ghaeini, R.; Fern, X.; Shahbazi, H.; and Tadepalli, P. 2019. Saliency Learning: Teaching the Model Where to Pay Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4016–4025. Minneapolis, Minnesota: Association for Computational Linguistics.

Glockner, M.; Habernal, I.; and Gurevych, I. 2020. Why do you think that? Exploring Faithful Sentence-Level Rationales Without Supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1080–1095. Online: Association for Computational Linguistics.

Hase, P.; Zhang, S.; Xie, H.; and Bansal, M. 2020. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4351–4367. Online: Association for Computational Linguistics.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205. Online: Association for Computational Linguistics.

Jacovi, A.; and Goldberg, Y. 2021. Aligning Faithful Interpretations with their Social Attribution. *Transactions of the Association for Computational Linguistics*, 9: 294–310.

Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; and Roth, D. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 252–262. New Orleans, Louisiana: Association for Computational Linguistics.

Kumar, S.; and Talukdar, P. 2020. NILE : Natural Language Inference with Faithful Natural Language Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8730–8742. Online: Association for Computational Linguistics.

Lehman, E.; DeYoung, J.; Barzilay, R.; and Wallace, B. C. 2019. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3705–3717. Minneapolis, Minnesota: Association for Computational Linguistics.

Lei, K.; Chen, D.; Li, Y.; Du, N.; Yang, M.; Fan, W.; and Shen, Y. 2018. Cooperative Denoising for Distantly Supervised Relation Extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, 426–436. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 107–117. Austin, Texas: Association for Computational Linguistics.

Li, S.; Zhao, S.; Cheng, B.; and Yang, H. 2018. An End-to-End Multi-task Learning Model for Fact Checking. In *Proceedings of the First Workshop on Fact Extraction and*

VERification (FEVER), 138–144. Brussels, Belgium: Association for Computational Linguistics.

Paranjape, B.; Joshi, M.; Thickstun, J.; Hajishirzi, H.; and Zettlemoyer, L. 2020. An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1938–1952. Online: Association for Computational Linguistics.

Petroni, F.; Piktus, A.; Fan, A.; Lewis, P.; Yazdani, M.; Cao, N. D.; Thorne, J.; Jernite, Y.; Plachouras, V.; Rocktäschel, T.; and Riedel, S. 2020. KILT: a Benchmark for Knowledge Intensive Language Tasks. In *arXiv:2009.02252*.

Pruthi, D.; Dhingra, B.; Neubig, G.; and Lipton, Z. C. 2020a. Weakly- and Semi-supervised Evidence Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3965–3970. Online: Association for Computational Linguistics.

Pruthi, D.; Dhingra, B.; Soares, L. B.; Collins, M.; Lipton, Z. C.; Neubig, G.; and Cohen, W. W. 2020b. Evaluating Explanations: How much do explanations from the teacher aid students? arXiv:2012.00893.

Pruthi, D.; Gupta, M.; Dhingra, B.; Neubig, G.; and Lipton, Z. C. 2020c. Learning to Deceive with Attention-Based Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4782–4793. Online: Association for Computational Linguistics.

Regulation, G. D. P. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *Official Journal of the European Union (OJ)*, 59(1-88): 294.

Ribeiro, M. T.; EDU, U.; Singh, S.; and Guestrin, C. 2016. Model-Agnostic Interpretability of Machine Learning. In *ICML Workshop on Human Interpretability in Machine Learning*.

Schuster, T.; Shah, D.; Yeo, Y. J. S.; Roberto Filizzola Ortiz, D.; Santus, E.; and Barzilay, R. 2019. Towards Debiasing Fact Verification Models. 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)*, 3419–3425. Hong Kong, China: Association for Computational Linguistics.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3319–3328. JMLR. org.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 809–819. New Orleans, Louisiana: Association for Computational Linguistics.

Treviso, M.; and Martins, A. F. T. 2020. The Explanation Game: Towards Prediction Explainability through Sparse Communication. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 107–118. Online: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30: 5998–6008.

Wang, Z.; and Culotta, A. 2020. Identifying Spurious Correlations for Robust Text Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3431–3440. Online: Association for Computational Linguistics.

Wiegreffe, S.; and Marasovic, A. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Wiegreffe, S.; Marasović, A.; and Smith, N. A. 2020. Measuring association between labels and free-text rationales. *arXiv preprint arXiv:2010.12762*.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4): 229–256.

Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D. I.; and Ravikumar, P. K. 2019. On the (In)fidelity and Sensitivity of Explanations. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; dÁlché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yu, M.; Chang, S.; Zhang, Y.; and Jaakkola, T. 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)*, 4094–4103. Hong Kong, China: Association for Computational Linguistics.

Zaidan, O. F.; Eisner, J.; and Piatko, C. 2008. Machine Learning with Annotator Rationales to Reduce Annotation Cost. In *Proceedings of the NIPS*2008 Workshop on Cost Sensitive Learning*.

Zhang, Y.; Marshall, I.; and Wallace, B. C. 2016. Rationale-Augmented Convolutional Neural Networks for Text Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 795–804. Austin, Texas: Association for Computational Linguistics.

Zhao, C.; Xiong, C.; Rosset, C.; Song, X.; Bennett, P.; and Tiwary, S. 2020. Transformer-XH: Multi-evidence Reasoning with Extra Hop Attention. In *The Eighth International Conference on Learning Representations (ICLR 2020)*.

Zhou, J.; Han, X.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; and Sun, M. 2019. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 892–901. Florence, Italy: Association for Computational Linguistics.