Rationalizing Transformer Predictions via End-To-End Differentiable Self-Training

Anonymous ACL submission

Abstract

We propose an end-to-end differentiable training paradigm for stable training of a rationalized transformer classifier. Our approach results in a single model that simultaneously classifies a sample and scores input tokens based on their relevance to the classification. To this end, we build on the widely-used three-player-game for training rationalized models, which typically relies on training a rationale selector, a classifier and a complement classifier. We simplify this approach by making a single model fulfill all three roles, leading to a more efficient training paradigm that is not susceptible to the common training instabilities that plague existing approaches. Further, we extend this paradigm to produce class-wise rationales while incorporating recent advances in parameterizing and regularizing the resulting rationales, thus leading to substantially improved and state-of-the-art alignment with human annotations without any explicit supervision.

1 Introduction

Neural networks are increasingly prevalent across a wide range of applications, driving significant advancements in fields such as natural language processing, computer vision, and beyond. Due to the black-box nature of these networks, this widespread use comes with an increased demand for interpretability (Lyu et al., 2024), as understanding the basis for the decisions made by these models is crucial for their reliable and ethical deployment. This need has become especially clear with the increasing use of notoriously uninterpretable large language models, which have the potential to quickly lose a user's trust after only few confidently incorrect predictions (Dhuliawala et al., 2023).

One possible mitigation is the use of encoder-only models, which lend themselves more readily to classical interpretability approaches designed for general neural network classifiers while still providing state-of-the-art performance due to continuous improvements in model structure (He et al., 2021, 2023) and training paradigms (Zhang et al., 2023).

While a variety of explainability methods exist, that usually assign scores to input tokens indicating their importance for a classification (Sun et al., 2021), these methods often suffer from several drawbacks, including high computational cost, difficult-to-interpret explanations, and potentially even unfaithful representations of the model's decision-making process. In this study, we close this gap by developing a rationalized transformer predictor that generates faithful and interpretable explanations in addition to its decisions within the same forward pass.

As a foundation for our approach, we build upon the existing and commonly used three-player game proposed by Yu et al. (2019). In this framework, a selector model chooses a subset of the input as rationale, while a predictor and a complement predictor model are trained to infer the correct label from either the tokens included in the rationale or the tokens not included in the rationale, respectively. The selector model is then trained to maximally aid the predictor in predicting the correct label while preventing the complement predictor from doing the same, thus ensuring that all tokens indicative of the correct label are included in the rationale.

While the general three-player game is sensible, the actual realizations that are proposed often have several limitations, including being not end-to-end differentiable due to a stochastic sampling process in the forward pass, showing interlocking dynamics that might prevent convergence to a suitable solution, and having no guarantee of providing a rationale that actually explains the prediction (compare Section 2.3 for a more detailed discussion).

For this reason, we propose a new take on this three-player game that is not susceptible to these drawbacks. We achieve this by making use of a single unified model that is trained as a standard classifier on the complete unaltered input, while simultaneously predicting class-wise importance scores for each input token in the same forward pass, which are then trained using self-training to mark spans that the model itself considers important for the specific class.

Our proposed rationalized transformer predictor (RTP) simplifies and enhances the common three-player structure in several ways, including 1) using only a single model to fulfill all three roles of the three-player game, thus enabling classification and rationale prediction in a single forward pass 2) training the rationales to explain the predictor, but avoiding training the predictor on the rationales, which ensures that the rationales faithfully explain the predictions 3) creating rationalized inputs in continuous fashion to enable fully differentiable training and avoid sampling 4) creating class-wise rationales and 5) using a parameterization that maximizes similarity with human rationale annotations.

We evaluate our method on two benchmarks for explainable AI and compare it with existing post-hoc explanation methods as well as methods leveraging standard multi-player procedures. We show that our method achieves state-of-the-art performance on both tasks, demonstrating previously unseen alignment with human rationales in combination with high rationale faithfulness.

2 Background

2.1 Post-Hoc Rationalization

Since neural networks are black-box models, the ever-increasing use of such model in research and industry has led to a strong demand for methods that reliably explain neural network classifications. To this end, a variety of approaches have been proposed, many of which are designed to create post-hoc explanations for an already trained classifier. These methods rely on a variety of mechanisms, including 1) making use of the models gradients at different inputs to obtain importance scores (Simonyan et al., 2013; Sundararajan et al., 2017) 2) quantifying the influence of individual input elements by observing the effect of input perturbations (Castro et al., 2009; Zeiler and Fergus, 2014; Zhou et al., 2014; Petsiuk et al., 2018) 3) fitting interpretable models to neural-network outputs (Ribeiro et al., 2016) 4) developing backpropagation-like procedures to propagate importance information from the model output to the input features (Zeiler and Fergus, 2014; Springenberg et al., 2015; Bach et al., 2015; Shrikumar et al., 2017; Chefer et al., 2021a,b) 5) performing input-optimization to create an altered input that only retains the information important for the classification (Brinner and Zarrieß, 2023).

2.2 Rationalized Classification

Due to the inherent difficulty of creating post-hoc explanations for classifiers that were never designed to be explainable, rationalized predictors have been proposed that are explicitly trained to perform the original task while simultaneously providing a rationale for the prediction in a single forward pass. Lei et al. (2016) were the first to propose a two-player game for textual inputs, involving a rationale selector model and a classifier model. The rationale selector assigns a probability to each input word, indicating its likelihood of belonging to the rationale, so that a discrete rationale can be sampled from this distribution. The classifier then uses only the rationale to make its classification, thus ensuring that the selected words were responsible for the classification. During training, the classifier is trained as usual to predict the correct label from a sampled rationale, while the rationale selector is trained to produce rationales that aid the classifier in making the correct predictions, ensuring that words indicative of the correct class are selected.

2.3 Common Issues of Rationalized Classifiers

While the general training paradigm of the two-player game is sensible, several issues affect the training, performance and faithfulness of the rationales:

1. **Stochastic Sampling**: Training requires stochastic sampling of rationales, meaning that gradients can only be estimated using methods like REINFORCE (Williams, 1992), which are generally less stable and slow to convergence.
2. **Class-Independent Rationales**: A single rationale is predicted regardless of the sample's class. In case a sample belongs to multiple classes, it is not possible to identify which part of the input is indicative of a specific class.
3. **Interlocking Dynamics**: Interlocking dynamics might lead to degenerate solutions, for example, if the rationale predictor adapts too quickly to the noisy rationales that are produced by the randomly initialized rationale selector or vice versa (Yu et al., 2021).

4. **Dominant Selector**: The training paradigm enforces rationales that persuade the classifier to predict a label that the predictor deemed correct, which does not necessarily correspond to faithful explanations of the actual reasoning process (Jacovi and Goldberg, 2021). In extreme cases, the rationale generator might simply encode the correct classification in the rationale (e.g., by selecting a specific kind of token), so that the classifier does not perform any significant reasoning itself.
5. **Mismatch with Human Annotations**: Often, rationales are most-useful if they resemble rationales provided by human annotators. Despite regularizers designed to enforce the selection of longer, consecutive spans of text, models often struggle to select spans that match human annotations, since overly strong regularization often overpowers the weak gradient signal created by REINFORCE, leading to degenerate solutions (e.g., selecting no tokens or all tokens).
6. **Degraded Classification Performance**: The actual classification performance often degrades compared to standard classifiers (Jacovi and Goldberg, 2021).

Several approaches have been proposed to modify or extend the two-player game to address these issues. Yu et al. (2019) extended this paradigm into a three-player game by introducing a complement predictor that is trained to predict the correct label from all words not included in the rationale. The rationale selector is then trained to prevent the complement predictor from identifying the correct class, thus ensuring that all words indicative of the correct class are selected as rationale, addressing issue 3 and (in part) issue 4. Chang et al. (2019) propose the CAR framework that uses two encoders and one decoder per class to generate class-wise and potentially counterfactual) rationales, solving issues 2, 3 and (in part) 4. They also use the straight-through gradient estimator (Bengio et al., 2013) instead of using REINFORCE, which addresses issue 1. Liu et al. (2023) make use of multiple generators to mitigate issue 3, while the A2R method (Yu et al., 2021) addresses the same issue by introducing a separate predictor that uses a soft selection of inputs instead of binary thresholding. To our knowledge, our proposed method for training a rationalized classifier is the only one to address all of the issues discussed above.

3 Method

We propose a new method for end-to-end differentiable training of a rationalized transformer predictor (RTP). In the following, plain letters (e.g., x) denote scalars, while bold letters (e.g., $\mathbf{x}$) denote vectors or tensors. We assume a text classification problem with label set $\mathcal{Y}$, and a training set consisting of texts $\mathbf{x}_0, ..., \mathbf{x}_n$ with corresponding ground truth vectors $\mathbf{y}_0, ..., \mathbf{y}_n$.

3.1 Concept

The RTP relies on a single model that, in one forward pass, produces both a classification output and class-wise importance scores for each token, denoting how indicative each token is of the respective class. The classification component is trained as a standard classifier, while the token-wise rationales are trained by creating altered inputs that only retain the important information for each individual class. The quality of these altered inputs (and therefore the quality of the rationales) is judged by the model itself by passing them through the model and observing its classification output. Through this end-to-end differentiable procedure, the rationales are optimized to faithfully explain the model predictions.

3.2 Model Structure

The basis of our method is a single model M, that, given an input text $\mathbf{x}$, simultaneously predicts class probabilities $\tilde{\mathbf{y}}$, as well as a mask tensor $\mathbf{m}$:

$$\tilde{\mathbf{y}}, \mathbf{m} = M(\mathbf{x}) \tag{1}$$

with the mask $\mathbf{m}$ being the rationale for the classification output $\tilde{\mathbf{y}}$. Notably, $\mathbf{m}$ consists of $|\mathcal{Y}|$ individual vectors $\mathbf{m}^0, ..., \mathbf{m}^{|\mathcal{Y}|}$ that constitute individual rationales for each class $c \in \mathcal{Y}$, with each $\mathbf{m}^c$ being a vector containing a mask value m_i^c in the range 0 to 1 for each input token x_i, indicating its influenc on the predicted likelihood of class c. In practice, the basis for classification output $\tilde{\mathbf{y}}$ will be the *CLS*-token embedding of the transformer classifier, while the mask values $\mathbf{m}$ will be calculated from the predicted outputs for each token.

3.3 Mask Parameterization

In this section, we will discuss the parameterization that transforms token-wise neural network outputs into a smooth mask. The RTP outputs a mask for each individual class, but since mask calculations for individual classes are independent of each other,

the start of this movie reminded me of parts from the movie stargate . people are looking around in an egyptian temple reading about some dangerous thing that is going to destroy earth in the future
.after a sort of confusing bit involving fake - looking cyborg things , the movie jumps into the future and the movie improves by leaps and bounds . the basic idea behind the movie is that every once
ina while (make that every 1000 years or so) an evil force comes to destroy earth . the things needed to defend against this menace are the four elements of nature plus the fifth element . the plot
inthis movie really is n ' t that important to the thing though . this movie has very good special effects , for the most part . the techno - ish music in the background fits the mood very well .
brucewillis is an illegal taxi - cab driver in a futuristic new york city . one day a lady draped with a few bandages drops down into his trunk . this movie is about what happens . the plot twists
areinteresting and the movie never fails to present the viewer with a variety of different locations . also there is a fair bit of action in the film , particularly towards the end . some characters
arejust plain strange including a highly - energetic deejay in drag . bruce willis does his normal job of blowing things away like he always does . the movie is definitely watchable and rarely slows
down. it is one of those sci - fi films where you ' ll be saying " cool " followed by a " what the hell ? ! ? ! " . i give the fifth element .

Figure 1: An exemplary output of the RTP for a positive review from the movie reviews dataset.

we will look at the mask $\mathbf{m}^c$ for a single class c, denoted as $\mathbf{m}$ for simplicity.

A simple mask parameterization would predict a logit l_i for each token x_i and define $m_i = \sigma(l_i)$. Even with regularizers that enforce smooth mask selections, this approach often fails to select long spans of text as rationales, which would be desirable for matching human annotations. For this reason, we opted for the mask parameterization proposed by Brinner and Zarrieß (2023) that explicitly enforces the prediction of longer spans of text as rationales by letting neighboring mask values influence each other.

In this parameterization, the model outputs two values w_i and σ_i for each word x_i. w_i is mainly responsible for determining the mask value of word x_i, while σ_i determines the influence of w_i on the mask values of neighboring words. Introducing regularizers to enforce large values for σ_i then leads to smooth masks. The mathematical formulation of the parameterization is as follows:

$$w_{i \to j} = w_i \cdot \exp\big(-\frac{d(i,j)^2}{\sigma_i}\big) \tag{2}$$

$$m_j = \text{sigmoid}(\sum_i w_{i \to j}) \tag{3}$$

Here, $d(i,j)$ denotes the distance between two words x_i and x_j and $w_{i \to j}$ is the influence of w_i on the mask value of word j. m_j is then calculated by applying the sigmoid to the sum of all influence values, resulting in the mask $\mathbf{m}$ for the specific class at hand. Predicting masks $\mathbf{m}^0, ..., \mathbf{m}^{|\mathcal{Y}|}$ for each class will simply be done by predicting individual outputs $\mathbf{w}^c$ and $\boldsymbol{\sigma}^c$ for each class c and performing the calculations independently.

3.4 Model Training

Given a sample $(\mathbf{x}, \mathbf{y})$, the classification capabilities of model M are trained like a standard neural network classifier by performing a prediction and applying a loss function like cross-entropy loss to the predicted output. In contrast to other rationalized models, our training paradigm therefore trains the classifier on the unaltered input, not on a masked version that might remove crucial information.

To train the rationale predictions (i.e., the masks $\mathbf{m}$), we use these masks to create two altered inputs $\mathbf{x}^c$ and $\overline{\mathbf{x}}^c$ for each ground-truth class c, with input $\mathbf{x}^c$ retaining all information that is indicative of class c according to mask $\mathbf{m}^c$, while $\overline{\mathbf{x}}^c$ is the complement input that removes all information specified by mask $\mathbf{m}^c$:

$$\mathbf{x}^c = \mathbf{m}^c \cdot \mathbf{x} + (1 - \mathbf{m}^c) \cdot b \tag{4}$$

$$\overline{\mathbf{x}}^c = (1 - \mathbf{m}^c) \cdot \mathbf{x} + \mathbf{m}^c \cdot b \tag{5}$$

Here, b denotes an uninformative background (e.g., *PAD*-token embeddings). Notably, the mask $\mathbf{m}$ is applied in continuous fashion to $\mathbf{x}$ and b, meaning that embeddings for words are linearly blended towards uninformative embeddings according to $\mathbf{m}$. In contrast to sampling of a discrete mask, this ensures full differentiability and was proven to have the desired effect of gradual removal of information by Brinner and Zarrieß (2023).

The rationalized inputs are then fed back into the same model M and are trained to allow it to predict the correct label from $\mathbf{x}^c$, but not from $\overline{\mathbf{x}}^c$, meaning that all information indicative of class c is contained in the rationale (thus enforcing rationale comprehensiveness). The loss formulations used are the following:

$$\mathcal{L}^c = \mathbf{CE}(M(\mathbf{x}^c), \mathbf{y}) \tag{6}$$

$$\mathcal{L}^{\overline{c}} = \text{relu}(M(\overline{\mathbf{x}}^c)[c] - \alpha) \tag{7}$$

where $\mathbf{CE}$ denotes the cross-entropy loss, $M(\mathbf{x})[c]$ denotes the predicted probability of class c and α is a hyperparameter ensuring that the model is not required to drive the probability of class c for input $\overline{\mathbf{x}}^c$ to 0, since driving it to a small value is sufficient. Importantly, the model M is only trained with respect to the first forward pass that produces the rationales, and is therefore not updated to improve classification on the altered inputs $\mathbf{x}^c$ and $\overline{\mathbf{x}}^c$. This ensures that the classification performance in not influenced, and that the rationales actually explain the classification instead of dictating it. The final optimization problem looks as follows:

$$\begin{aligned} \arg\min_{M} \;\; & \mathbf{CE}(M(\mathbf{x}),\mathbf{y}) + \sum_{c \in \mathbf{y}} \big[\mathcal{L}^c + \mathcal{L}^{\overline{c}}\big] \\ & + \Omega_\lambda + \Omega_\sigma \end{aligned} \tag{8}$$

where $c \in \mathbf{y}$ indicates summing over all ground-truth labels and Ω_λ and Ω_σ denote regularizers that enforce sparsity and smoothness of the rationales, respectively. Details on these regularizers and further details on the general objective are available in Appendix A.3.

3.5 Advantages

Our proposed scheme solves all issues discussed in Section 2.3, and is (to our knowledge) the only method to do so. The main advantages lie in the fully differentiable formulation that does not require sampling and gradient approximation, the fact that class-wise rationales are created, and especially in the fact that the classifier is trained on the unaltered inputs instead of on the rationalized variants. This last point ensures that the rationales do not dictate the classification result, but instead explain the actual classification made by the classifier. It also means that no degradation is classification performance is to be expected.

4 Experiments

We evaluate our method with regards to matching human evidence annotations for text classifications, as well as with regards to the faithfulness of the explanations with regards to the classifier. For details regarding the model and the training and prediction procedures, see Appendix A.

4.1 Datasets

In our evaluation, we use two text classification datasets with span-level evidence annotations, each posing different challenges. The first is the movie review dataset (Zaidan et al., 2007), containing 2000 reviews with sentiment labels (*positive* or *negative*) and span-level evidence annotations. For this dataset, DeYoung et al. (2020) provided more comprehensive rationales for the test split, which we use in our evaluation. Since class labels are mutually exclusive, this dataset allows models to perform optimally even without class-wise rationales. Additionally, this dataset enables optimal assessment of agreement between predicted rationales and human annotations, since the simplicity of the classification task eliminates the lack of understanding of the inputs as a cause for mismatches.

As for a more challenging classification task, we use the INAS dataset (Brinner et al., 2022), consisting of 954 scientific paper titles and abstracts from the domain of invasion biology together with labels indicating which hypothesis (from a set of 10 common hypotheses in the field) is addressed in each paper. In a subsequent study, Brinner et al. (2024) provided span-level evidence annotations for 750 of the samples. Since some samples have multiple correct labels, optimal performance on this dataset requires class-wise rationalization. Additionally, the more challenging nature of the classification task can highlight degraded classification performance of rationalized models.

4.2 Evaluation Metrics

We evaluate the consistency with human annotations on token-level and span-level as done in (Brinner et al., 2024), and evaluate the faithfulness of rationales with respect to the classifier as done in (Brinner and Zarrieß, 2023).

Token-Level Evaluation To evaluate agreement with human rationales at the token level, we use the area under the precision-recall curve (*AUC-PR*). We also assess the token-level F1 score (*Token-F1*), which requires binary predictions. This is done by selecting the highest-scoring p percent of tokens as positive predictions, calculating the standard F1 score, and averaging over 19 values of p (5, 10, ..., 95). For a better absolute assessment of prediction quality, we use the discrete token-level F1 score (*D-Token-F1*), where the top k tokens (matching the number in the ground-truth annotation) are treated as the rationale and evaluated with the F1 score.

Span-Level Evaluation We also evaluate the quality of predicted spans of text, defined as consecutive words selected as part of the rationale after binary thresholding. The span-level IoU-F1 score (*IoU-F1*) is calculated by determining spans in both the binary rationale prediction and the ground-truth annotation, calculating the IoU for all span pairs, and selecting the maximum IoU for each predicted and annotated span. IoU-precision and IoU-recall are then defined as the averages of these maximum IoU values for predicted and ground-truth spans, respectively. The holistic IoU-F1 score is then obtained by averaging over the same 19 discrete token selections used for the token-level F1 score. The discrete IoU-F1 score (*D-IoU-F1*) is again calculated by selecting the top-scoring tokens to match the number in the ground-truth annotation.

Faithfulness Evaluation We evaluate faithfulness using scores for sufficiency and comprehensiveness of the predicted rationales. The sufficiency score measures the model's ability to predict the correct label using only the highest-scoring words

Method	Clf-F1	AUC-PR	Token-F1	D-Token-F1	IoU-F1	D-IoU-F1	Suff. ↓	Comp.↑	Perf.
Random	-	0.220	0.255	0.222	0.067	0.003	0.194	0.191	0.289
Supervised	0.730	0.557	0.406	0.509	0.231	0.257	0.005	0.396	1.028
MaRC	**0.748**	0.366	0.336	0.351	0.219	0.178	-0.002	0.396	0.953
Occlusion	**0.748**	0.307	0.277	0.294	0.145	0.071	0.020	0.315	0.717
Int. Grads	**0.748**	0.315	0.302	0.318	0.087	0.013	-0.017	0.465	0.871
LIME	**0.748**	0.272	0.280	0.273	0.082	0.007	0.039	0.357	0.680
Shapley	**0.748**	0.309	0.301	0.320	0.084	0.009	**-0.083**	0.515	0.983
L2E-MaRC	**0.748**	0.431	**0.359**	0.402	0.174	0.131	0.020	0.427	0.940
2-Player	0.675	0.272	0.286	0.270	0.085	0.007	-0.050	0.303	0.724
3-Player	0.722	0.287	0.296	0.286	0.080	0.004	0.023	0.403	0.756
CAR	-	0.314	0.281	0.280	0.184	0.133	-	-	-
A2R	0.686	0.268	0.287	0.264	0.084	0.008	0.122	0.282	0.531
A2R-Noise	0.618	0.271	0.285	0.262	0.087	0.011	0.072	0.198	0.498
RTP	0.710	**0.436**	**0.359**	**0.415**	**0.220**	**0.203**	0.066	**0.565**	**1.078**

Table 1: Results on the INAS dataset, divided into groups of standard-baselines, post-hoc explainability methods and rationalized neural networks. Best scores per metric are bold, second best are underlined.

in the rationale. A lower sufficiency score indicates that fewer tokens are needed for a correct prediction, thus indicating a more faithful rationale:

$$\text{sufficiency}(x, r) = \frac{1}{19} \sum_{i=1}^{19} M(x) - M(r_i) \quad (9)$$

The comprehensiveness score is higher if removing the highest-scoring words according to the rationale quickly degrades the model's predictions, again indicating faithful rationales:

$$\text{comp}(x, r) = \frac{1}{19} \sum_{i=1}^{19} M(x) - M(x \backslash r_i) \quad (10)$$

In these equations, x denotes the input sample, r_i denotes the $(i \cdot 5)\%$ of input tokens with the highest scores according to the rationale, $x \backslash r_i$ denotes the input x with the tokens from r_i removed, and $M(x)$ denotes the probability that model M assigns to the correct class given input x. To avoid relying on a single threshold, these scores are calculated by summing over different percentages of rationale tokens used or removed, respectively.

Overall Performance Ideally, a model should produce rationales that both agree with human rationales and demonstrate faithfulness. We therefore provide an overall performance score (*Perf.*) that sums over the Token-F1, IoU-F1, comprehensiveness and negative sufficiency scores, thus assessing agreement and faithfulness comprehensively.

4.3 Baseline Methods

We compare our rationalized transformer predictor (RTP) against other rationalized classifiers, which are a two-player game as proposed by Lei et al. (2016), a three-player structure with complement predictor (Yu et al., 2019), the CAR framework for class-wise rationale generation (Chang et al., 2019), and the A2R method (Yu et al., 2021) as well as an extension to it using noise injection (Storek et al., 2023). We also compare post-hoc explainability methods that are applied to a standard classifier, which includes MaRC (Brinner and Zarrieß, 2023), Occlusion (Zeiler and Fergus, 2014), Integrated Gradients (Sundararajan et al., 2017), LIME (Ribeiro et al., 2016), Shapley value sampling (Castro et al., 2009), as well as a neural network predictor trained on MaRC rationales (L2E-MaRC, Situ et al. (2021)). Finally, we report results for a supervised model trained on rationale annotations and a random predictor as additional baselines. For a more detailed overview, see Appendix A.1.

5 Results

The results for the evaluation on the movie review dataset and the INAS dataset are displayed in Table 2 and Table 1, respectively. Exemplary predictions are displayed in Figure 1, with further examples being included in Appendix B.

5.1 Classification Performance

On the INAS dataset, our RTP method ranks among the best-performing rationalized neural networks, although all rationalized models show slightly worse classification performance compared to a standard classifier (used by the post-hoc methods). We do not attribute this to a generally decreased classification ability of the rationalized classifiers, since we selected the best-performing version of each model with respect to rationale-predictions

Method	Clf-F1	AUC-PR	Token-F1	D-Token-F1	IoU-F1	D-IoU-F1	Suff. ↓	Comp.↑	Perf.
Random	-	0.316	0.326	0.312	0.061	0.002	0.227	0.238	0.398
Supervised	0.980	0.670	0.514	0.626	0.144	0.169	0.001	0.638	1.295
MaRC	0.965	0.428	0.404	0.423	0.181	0.118	0.036	0.478	1.027
Occlusion	0.965	0.409	0.367	0.377	0.151	0.079	-0.021	0.569	1.108
Int. Grads	0.965	0.376	0.358	0.371	0.067	0.009	0.049	0.484	0.860
LIME	0.965	0.379	0.361	0.369	0.076	0.014	0.005	0.603	1.035
Shapley	0.965	0.442	0.390	0.426	0.082	0.020	**-0.029**	**0.827**	1.328
L2E-MaRC	0.965	**0.565**	**0.460**	**0.534**	0.126	0.104	-0.016	0.652	1.254
2-Player	0.930	0.516	0.449	0.508	0.113	0.066	-0.024	0.210	0.796
3-Player	0.955	0.458	0.422	0.465	0.089	0.023	0.003	0.354	0.862
CAR	-	0.384	0.364	0.376	0.078	0.013	-	-	-
A2R	0.955	0.474	0.433	0.486	0.111	0.046	0.109	0.320	0.755
A2R-Noise	0.950	0.483	0.440	0.492	0.107	0.044	0.005	0.338	0.880
RTP	**0.975**	0.558	0.458	0.527	**0.192**	**0.177**	-0.006	0.803	**1.459**

Table 2: Results on the movie reviews dataset, divided into groups of standard-baselines, post-hoc explainability methods and rationalized neural networks. Best scores per metric are bold, second best are underlined.

on the validation set, not with respect to classification performance. Additionally, this task shows a high variance between training runs (Brinner et al., 2022). Results on the movie review dataset are similar, with most rationalized classifiers performing slightly worse than the standard classifier. Notably, the RTP even outperforms the baseline, which we again attribute to variance between training runs.

5.2 Token-Level Performance

For token-level rationale evaluations on the INAS dataset, our RTP method outperforms others in AUC-PR, token-F1, and discrete token-F1 scores. Only the L2E-Marc method, which is a neural network trained to predict rationales created by the MaRC method, comes close to or matches the RTP. A discussion of the superior performance of these exact two methods is done in Section 6. Especially the ability to predict class-wise rationales benefits the RTP and the post-hoc methods on this dataset compared to the other rationalized neural networks, since only the CAR method possesses this ability among them. This can also be seen by noticing the much better performance of these rationalized networks on the movie review dataset, since on this task the ability to predict class-wise rationales is irrelevant, thus reducing the gap to the RTP and making them surpass most post-hoc methods. Generally, the supervised baseline outperforms all methods with regards to token-level predictions, which is to be expected considering that the weakly supervised methods were never explicitly told what to predict. However, the RTP comes rather close to the supervised baseline in many metrics, indicating that in the absence of ground-truth rationales for training, using the weakly supervised scheme can be an effective alternative.

5.3 Span-Level Performance

Our RTP method is consistently the best performing method with regards to the IoU-F1 and the discrete IoU-F1 scores. Compared to the other rationalized methods this is to be expected, since the RTP has been explicitly designed to extract longer spans of text as rationales. In contrast, other rationalized predictors often rely on a total variation regularizer in their optimization objective, which we found to be ineffective since increasing it's strength quickly leads to degenerate solutions with either all or none of the words being selected. This highlights the importance of using the MaRC mask parameterization that reliably leads to the desired results. Notably, the RTP comes close to the supervised method on the INAS dataset without any supervision regarding the usual form of human annotations. On the movie review dataset, the RTP even outperforms the supervised method due to the rationales from the test set being more extensive, thus causing a mismatch between training and test data distributions. This shows, that even if slightly inaccurate training data is available for a given task, using a weakly supervised method instead might be preferable.

5.4 Faithfulness Results

Our RTP method achieves competitive sufficiency scores and state-of-the-art comprehensiveness scores. We found that assigning high scores to few important words distributed throughout the whole input is a great strategy for achieving high sufficiency, since the model can quickly recognize the correct label from these few highly indicative words. Our RTP model still performs well despite being explicitly discouraged from pursuing this strategy, indicating that our optimization objective is reasonable for generating faithful rationales.

For comprehensiveness, the RTP attains state-of-the-art results on the INAS dataset and nearly matches the Shapley value sampling method on the movie reviews dataset, with both methods outperforming all other contenders by a large margin. Good faithfulness scores for Shapley value sampling are expected, though, since its objective for scoring input tokens aligns closely with the evaluation measures for faithfulness.

Overall, our RTP method compares favourably to other rationalized neural networks, since it optimizes its rationales to actually explain the classification, while other methods suffer from issues like a dominant predictor that already dictates a specific label, and additionally train the predictor on the rationales, which leads to a constant mismatch between the current predictor and the predictor that the rationales have been trained to explain.

Another important insight is, that post-hoc explanation methods do not offer an advantage over the rationales generated by the RTP. Considering, that post-hoc explainers outperform other rationalized networks with respect to faithfulness of the explanations, our method is the first all-in-one method that offers both predictions and rationales with state-of-the-art faithfulness in a single forward pass.

5.5 Overall Performance

As discussed, the RTP achieves state-of-the-art results in agreement with human rationales and rationale faithfulness, resulting in dominant scores for overall performance (*Perf.*) on both tasks. In comparison, other rationalized neural networks fall significantly short, with only few post-hoc methods coming somewhat close. These methods have the downside of a substantially higher computational cost in producing a rationale, with, for example, MaRC and Shapley value sampling requiring hundreds of forward passes to create a single rationale.

6 Discussion

The RTP model demonstrated strong performance across all evaluated metrics. Comparing it specifically to the MaRC method, it outperformed it in every metric related to measuring agreement with human annotations and most faithfulness metrics. This is notable since the RTP can be seen as a neural network parameterized version of the MaRC approach, which originally optimized mask parameters for each sample individually instead of training a neural network to directly predict them from the input. Another well-performing method, especially with regards to token-level evaluation, is the L2E-MaRC method. The L2E framework (Situ et al., 2021) trains a neural network on pre-calculated rationales created by a post-hoc explainer. Even though it only saw rationales produced by the MaRC method, it manages to outperform it on all metrics measuring token-level agreement with human rationales. These two results indicate, that training to explain many different samples leads to better generalization, which we attribute to reduced overfitting to one specific input. This effect is crucial for the RTP, since it performs input optimization with respect to specific neural network outputs, which has been shown to generally lead to unexpected and uninterpretable artifacts (Simonyan et al., 2013). The MaRC method successfully mitigated this issue by combining constrained optimization with heavy regularization, but artifacts (i.e., unexpected spans included in the rationale) are still to be expected. In the case of the RTP, training on many samples further reduces this issue, since these unwanted gradient signals will generally not match between different samples, so that the neural network mainly adapts to the wanted signal that is consistent within larger parts of the training set, and that corresponds to features that are generally indicative of the respective class.

7 Conclusion

We presented a new method for training a rationalized transformer predictor and demonstrated its strong performance on two natural language processing benchmarks. Since our proposed training scheme is not invasive to the general training process and does not produce significant overhead during prediction, we believe that this approach has the potential to facilitate wider adoption and availability of rationalized predictors. Given that transformers are widely used in other modalities (Dosovitskiy et al., 2021; Verma and Berger, 2021), we hypothesise that our approach can be extended to these modalities and potentially lead to results of similar quality.

8 Limitations

While our method for rationalization generally does not interfere with the training of the prediction module and does not produce notable overhead during prediction, it nevertheless increases the computational cost of model training due to a second forward pass through the model, as well as through more training epochs being required due to slower convergence of rationale training compared to the classification component.

Additionally, the exact form of the produced rationales depends on the models inner working, so that generally a high overlap with human rationales is not guaranteed in cases where the model's reasoning and human reasoning differ.

Finally, while having access to word-level rationale scores is generally helpful, this does not equate to a complete description of the model's inner workings and the actual reasoning process, which most likely is impossible to represent in such a simple form.

References

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.

Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv*, abs/1308.3432.

Marc Brinner, Tina Heger, and Sina Zarriess. 2022. Linking a hypothesis network from the domain of invasion biology to a corpus of scientific abstracts: The INAS dataset. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 32–42, Online. Association for Computational Linguistics.

Marc Brinner and Sina Zarrieß. 2023. Model interpretability and rationale extraction by input mask optimization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13722–13744, Toronto, Canada. Association for Computational Linguistics.

Marc Brinner, Sina Zarrieß, and Tina Heger. 2024. Weakly supervised claim localization in scientific abstracts. In *Robust Argumentation Machines (RATIO 2024)*, Bielefeld, Germany. Springer.

Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730. Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2019. A game theoretic approach to class-wise selective rationalization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hila Chefer, Shir Gur, and Lior Wolf. 2021a. Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 387–396.

Hila Chefer, Shir Gur, and Lior Wolf. 2021b. Transformer interpretability beyond attention visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

Michael G. Cripps, Graeme W. Bourdôt, David J. Saville, Hariet L. Hinz, Simon V. Fowler, and Grant R. Edwards. 2011. Influence of insects and fungal pathogens on individual and population parameters of cirsium arvense in its native and introduced ranges. *Biological Invasions*, 13(12):2739–2754.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. A diachronic perspective on user trust in AI under uncertainty. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5567–5580, Singapore.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Catherine S. Jarnevich, Thomas J. Stohlgren, David Barnett, and John Kartesz. 2006. Filling in the gaps: modelling native species richness and invasions using spatially incomplete data. *Diversity and Distributions*, 12(5):511–520.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *Preprint*, arXiv:2009.07896.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Xinyang Li, YuanKai Zhang, and Yang Qiu. 2023. MGR: Multi-generator based rationalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12771–12787, Toronto, Canada. Association for Computational Linguistics.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2).

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327. Association for Computational Linguistics.

Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Xuelin Situ, Ingrid Zukerman, Cecile Paris, Sameen Maruf, and Gholamreza Haffari. 2021. Learning to explain: Generating stable explanations fast. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5340–5355, Online. Association for Computational Linguistics.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2015. Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Adam Storek, Melanie Subbiah, and Kathleen McKeown. 2023. Unsupervised selective rationalization with noise injection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12647–12659, Toronto, Canada. Association for Computational Linguistics.

Xiaofei Sun, Diyi Yang, Xiaoya Li, Tianwei Zhang, Yuxian Meng, Han Qiu, Guoyin Wang, Eduard Hovy, and Jiwei Li. 2021. Interpreting deep learning models in natural language processing: A review. *Preprint*, arXiv:2110.10470.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Prateek Verma and Jonathan Berger. 2021. Audio transformers:transformer architectures for large scale audio understanding. adieu convolutions. *Preprint*, arXiv:2105.00335.

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

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Mo Yu, Yang Zhang, Shiyu Chang, and Tommi Jaakkola. 2021. Understanding interlocking dynamics of cooperative rationalization. In *Advances in Neural Information Processing Systems*, volume 34, pages 12822–12835. Curran Associates, Inc.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.

Zhen-Ru Zhang, Chuanqi Tan, Songfang Huang, and Fei Huang. 2023. Veco 2.0: Cross-lingual language model pre-training with multi-granularity contrastive learning. *Preprint*, arXiv:2304.08205.

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2014. Object detectors emerge in deep scene cnns. *CoRR*, abs/1412.6856.

A Experimental Details

The code for our experiments is available at *Link will be added in final version of the paper.*

A.1 Baseline Methods

We evaluate our rationalized transformer predictor against a variety of baseline methods that pursuit different strategies for rationalizing predictions. This section provides a general overview, with many model and training details being discussed in Appendix A.2. The first group are rationalized neural networks, that learn to create rationales from sample-level labels alone. We evaluate the performance of the following methods:

- **2-Player**: A two-player structure using a rationale extractor and a predictor as proposed by Lei et al. (2016). We used an own implementation, since the original work did not use transformers.
- **3-Player**: A three-player structure using a rationale extractor, a predictor and a complement predictor as proposed by Yu et al. (2019). We used an own implementation, since the original work did not use transformers.
- **CAR**: The CAR framework for creating class-wise rationales (Chang et al., 2019). We use an own implementation, since the original work did not use transformers. Additionally, we use more extensive parameter sharing, as the original work use a separate rationale predictor for each class, which is impracticable especially for the 10-class classification problem on the INAS dataset. Therefore, a single BERT model predicts rationales for each class at the same time, while a second BERT model acts as the single predictor.
- **A2R**: The A2R framework as proposed by (Yu et al., 2021). We use the implementation of (Storek et al., 2023), who created an implementation relying on BERT models, which, according to their evaluation, outperformed the original implementation that relies on GRUs.
- **A2R-Noise**: The A2R framework with additional noise injection as proposed by (Storek et al., 2023). We use the implementation provided by the original study.

We also evaluated a variety of post-hoc explanation methods:

- **MaRC**: The MaRC method as proposed by (Brinner and Zarrieß, 2023). We use the updated weight regularizer proposed by (Brinner et al., 2024).
- **Occlusion**: The occlusion method as proposed by Zeiler and Fergus (2014). We chose to mask slightly larger spans of 5 tokens as this produced smoother masks which resulted in higher IoU F1 scores. We use the implementation by Kokhlikyan et al. (2020).
- **Int. Grads**: The integrated gradients method (Sundararajan et al., 2017). We use the implementation by (Kokhlikyan et al., 2020).
- **LIME**: The LIME method (Ribeiro et al., 2016). We train a linear classifier on scores from 50 function evaluations. In each evaluation, $5 - 13\%$ of tokens are selected and the thee tokens starting from the chosen token are removed as input perturbation. We use the implementation by (Kokhlikyan et al., 2020).
- **Shapley**: Shapley value sampling (Castro et al., 2009). We perform 25 feature permutations per sample, and use the implementation by (Kokhlikyan et al., 2020).
- **L2E-MaRC**: The L2E framework (Situ et al., 2021). We use the rationales created by the MaRC method on the training samples. We discretize the rationales into 5 bins and train

a classifier on this dataset. For scoring, we predict the bin-probabilities for each word, multiply them by the bin-means and sum over the resulting values to get a single, continuous score for each word.

We also evaluate two further baselines: A random baseline that predicts random scores for each input token, and a supervised method that is trained to perform a binary prediction on each individual token from the input.

A.2 Model and Training Details

Base Models For the movie review experiment, we use bert-base-uncased (Devlin et al., 2019) as base model to stay consistent with previous work and the be able to use existing code bases to ensure implementational accuracy. For the INAS dataset, we use PubMedBERT-base-uncased (Gu et al., 2021), since it shows strong performance on the standard classification task for this dataset (Brinner et al., 2022). These base classifiers are used for all baseline methods, and for all parts of the pipelines (encoder, predictor, base classifier, etc.).

Input Processing During training, samples that exceed the 510 token limit for BERT models were split into multiple segments, and one segment was chosen randomly for this model update. For the evaluation, we again split each sample into smaller parts that adhere to the token limit and that overlap for 100 tokens. Scores were predicted for each split separately and linearly blended afterwards.

Model Selection During training, evaluations on the validation set were performed after each epoch, and the best-performing version of the model was selected for testing. On the INAS dataset, the agreement of the predicted rationales with the human annotations was evaluated after each epoch, and the mean of all five scores (AUC-PR, Token-F1, D-Token-F1, IoU-F1, D-IoU-F1) was taken as performance indicator. On the movie dataset, this procedure was not possible, since the data distribution of the validation and test samples is different, meaning that validation results are not a good indicator for performance on the test set. Especially the span-level evaluation scores were unsuitable, since much shorter spans were annotated on the validation set. We chose to use the AUC-PR as performance measure, since it still indicates the models ability to generally recognize useful words.

A.3 RTP Objective Details

We use two regularizers in the optimization objective for our rationalized transformer predictor. The first is a sparsity regularizer that ensures that only a subset of tokens is selected as rationale:

$$
\begin{aligned}
\Omega_\lambda = & \sum_{c \in \mathbf{y}} \alpha_1 \cdot \mathrm{mean}(\mathbf{m}^c)^2 + \alpha_2 \cdot \mathrm{mean}(\mathbf{m}^c) \\
& + \sum_{c \notin \mathbf{y}} \alpha_3 \cdot \mathrm{mean}(\mathbf{m}^c)^2 + \alpha_4 \cdot \mathrm{mean}(\mathbf{m}^c)
\end{aligned}
$$

In summary, we perform L1 and L2 regularization on the mask means for the masks of the ground truth classes and non-ground-truth classes. In our experiments, we used $\alpha_1 = 0.2$ and $\alpha_3 = 0.05$, meaning that regularization for masks of incorrect classes is weaker. We chose this setting, since for incorrect classes there is no signal that forces words the be unmasked, so that less strong regularization is required. The L1 parameters are set to rather low values of $\alpha_2 = \alpha_4 = 0.001$.

The smoothness regularizer has the following form:

$$
\Omega_\sigma = \beta_1 \cdot \sum_{c \in \mathcal{Y}} \mathrm{mean}((\boldsymbol{\sigma}^c - \beta_2)^2)
$$

Here, the inner subtraction is meant to be element-wise, so that we regularize each individual sigma value towards a value of β_2. The actual hyperparameters used are $\beta_1 = 0.02$ and $\beta_2 = 3$.

Finally, we use individual weights for each major component of the optimization objective (Equation 8):

$$
\begin{aligned}
\arg\min_M \quad & \gamma_1 \cdot \mathbf{CE}(M(\mathbf{x}), \mathbf{y}) \\
& + \sum_{c \in \mathbf{y}} \left[\gamma_2 \cdot \mathcal{L}^c + \gamma_3 \cdot \mathcal{L}^{\overline{c}} \right] \\
& + \gamma_4 \cdot \Omega_\lambda \\
& + \gamma_5 \cdot \Omega_\sigma
\end{aligned}
$$

These values are set to $\gamma_1 = 2$, $\gamma_2 = 5$, $\gamma_3 = 10$, $\gamma_4 = 6$ and $\gamma_5 = 6$.

A.4 Evaluation Post-Processing

In the INAS dataset, rationales do not cross sentence boundaries. For that reason, we opted to employ a post-processing step that uses SciSpacy (Neumann et al., 2019) to split each abstract into sentences, and set the rationale score of the last token in each sentence (that corresponds to punctuation) to 0. This is done for all methods and generally lead to a slight improvement in agreement scores.

B Examples

B.1 Movie Reviews Dataset

hedwig (john cameron mitchell) was born a boy named hansel in east berlin . as a teen seeking his " other half , " he reluctantly agrees to a sex change operation in order to marry american g . i . luther(maurice dean wint) . the operation , performed by a hack surgeon , is botched , and the " angry inch " is all that ' s left . now a " she " , hedwig comes to america , is abandoned by luther ,forms a rock band and falls for her 17 - year old lover / prot ? g ? , tommy , only to be rejected by him later , too . she and her band , the angry inch , shadows the now - famous tommy gnosis acrossthe us (for revenge ?) , but hedwig is really in search of her lost other half in " hedwig and the angry inch . " first time helmer john cameron mitchell , along with composer / lyricist stephentrask , created and starred in their acclaimed off - broadway production that has become the movie . and quite a movie it is in its eclectic variety of songs , outrageous costumes , sets and makeupand , especially , a riveting performance by mitchell as the title character . mitchell and trask have reinvented the movie musical and couple it with the underlying story of just whom hedwig isand what she is looking for . i am , by far , not a big fan of musicals . sure , there are exceptions , like bob fosse ' s " all that jazz " and " cabaret , " but , for the most part , they are justnot my cup of tea . " hedwig and the angry inch " is an exception , though , with its combination of humor , wit and a collection of tunes that covers musical styles ranging from " the rocky horrorpicture show " and meatloaf to david bowie to the sex pistols . the original songs , by stephen trask (also appearing as one of the members of the band the angry inch) , are full of energy andvariety and , even though it ' s not my kind of music , i found every one entertaining and fun . the audience i saw " hedwig " with thought so , too . the main attraction to this one - man / woman showis the presence of its star . john cameron mitchell gives a solid , sometimes fun , sometimes angry performance as a person searching for self - enlightenment and love . as a young boy growing up ineast berlin , hansel is abused by his g . i . father and raised by his german mother in a tiny flat so small that " mother would make me play in the oven " where he listened to pop music on armed forcesradio . later , as a young man , he meets luther , another g . i . and is swept off of his feet . the ensuing angry inch incident comes soon after . flash forward to a trailer park in junction city, kansas , and luther is leaving hedwig for another boy . frustrated and broke , she takes on baby - sitting and the odd " job " to make ends meet . she also forms a band with four korean housewivesand the musical talent of hedwig is born . she meets , falls for and loses young tommy , who steals her songs and goes off to become a rock sensation . jealous and angry , hedwig and her newband begins a campaign to shadow tommy ' s tours and , with the help of her manager , phyllis stein (andrea martin) , is trying get a law suit going against the star for stealing her songs . hedwigand the angry inch get gigs , not coincidentally , at a chain of seafood restaurants that just happen to be next to the forums where tommy gnosis is playing . things finally come to a head , so tospeak , in new york city . the popularity of the off - broadway musical and its offshoots have garnered a ready - made audience base for " hedwig . " the wit , humor , music and search for identity hasgreat appeal to young adults , but the charismatic presence of mitchell makes this a cut above what it could have been . it is this one - man / woman show that casts its spotlight on its " internationallyignored " rock star and mitchell is outstanding in the role . there is n ' t a lot going on with other characters , though there are amusing little sidebars , like hedwig ' s backup singer/ lover , yitzak (miriam shor) , deciding to break away from the band to join a polynesian road show of " rent " as a puerto rican drag queen . the low budget that the moviemakers have for theproduction belies the quality of the film . attention to details - hedwig ' s costumes and outrageous " cabaret " - like makeup ; the seedy trailer park setting ; and , the kitschy seafood restaurants- are loads of fun to watch and lend the appropriate air to the proceeds , all on what has to be a beer budget . " hedwig and the angry inch " may not be for everybody , but the energy of theeffort , the songs , the imaginative sets and costumes and a fast steady pace make it a pleasure to watch . if you ' re a fan of contemporary , edgy music , it is an even bigger draw . i give it a b+ .

Figure 2: An exemplary output of the RTP for a positive review from the movie reviews dataset. Green text indicates the ground-truth annotations.

writing a screenplay for a thriller is hard . harder than pouring concrete under the texas sun . harder than building a bridge over troubled waters . and incidentally , a whole heck of a lot harder thanwriting a movie review . thrillers are all variations on a theme . you have a smart , resourceful , and powerful bad guy , who has a goal he has to meet . you have a noble and brave good guy , whohas to protect the innocent , kill the bad guy , and not get killed himself in the process . the trick of thriller writing is doing all of this in an interesting and novel manner . this simple formulacan lead to classic movies like north by northwest , high noon , or silence of the lambs , or big summer blockbusters like men in black , the fugitive , or air force one , or it can lead to utterdreck like masterminds , event horizon , kull the conqueror . . . . is anyone else getting depressed here ? point is , it ' s not enough to follow the formula . you ' ve got to throw in somethingextra , something good and new and better than the last version . something to surprise and move all of us people who buy the tickets and the popcorn and the happy meals . this is a hard thingto do , but it is absolutely necessary in every way . without that something extra - - whether it ' s a great plot or a well - written screenplay , or great special effects or great locations or greatcasting or great performances or great big hungry dinosaurs - - the movie fails . that ' s why the jackal , with all its starpower , with all its budget , with all its hype , gets a big fat f . brucewillis is the bad guy , the jackal , a legendary killer for hire . richard gere is the good guy , a former ira assassin with a vendetta against the jackal . the jackal is trying to kill someone .gere is trying to stop him . will gere be able to stop the assassination in time and kill the jackal ? (i ' ll give you three guesses , and the first two do n ' t count .) there are no surprises awaitingthe audience in the jackal , no moment when you say to yourself , " i wonder what happens next ? " the script for the jackal is n ' t ripped straight from today ' s headlines . it ' s ripped off, straight from an episode of millennium . throughout the movie , we learn what the jackal ' s plans are and how he intends to accomplish them . no surprise . the fun of a movie like this should comefrom richard gere figuring out what the jackal ' s plan is and developing a clever plan to foil the bad guy . instead , we get two (count ' em , two) scenes where gere is sitting in an fbi conferenceroom somewhere and instantly divines the jackal ' s plan just as if he ' s frank black (or more likely , just as if he ' s been handed a copy of the script) . and we never get more than a superficialclue as to why gere has had this flash of insight . it ' s like gere ' s character is psychic , but neither he nor the fbi (or the screenwriters) seem to know it . and just like in millennium, the bad guy has an overwhelming need to go after the people the good guy cares about , whether or not they are important to what he ' s trying to do or not . what ' s more , in the last halfof the movie , the jackal , supposedly a super - smart professional terrorist who never makes a mistake , comes down with a major case of the stupids . as for the performances . . . bruce willis managesto get through the whole movie without a wisecrack , which is a major achievement , but not enough reason to see the movie . his disguises are good , but not as good or as interesting as val kilmer' s in the saint . richard gere is made to talk the entire movie in an irish accent , which detracts from his otherwise lifeless and dull performance . sidney poitier is probably the most disappointingelement in a overwhelmingly disappointing movie - - not that his performance is bad or anything , it ' s not , but it is sad that hollywood wo n ' t use this talented actor in any part otherthan an fbi agent (shoot to kill , sneakers) . writing a good plot and a good screenplay , like i said , is hard , but it can be done . it was n ' t done here . it is our job as consumers to rewardgood screenplays and to denounce bad and uninteresting ones . do not go see this movie . you ' ll only encourage the producers to make more just like it . instead , stay home and rent day of thejackal , or in the line of fire , or a fire safety video , for crying out loud . anything other than the jackal , which lives up to its name by gnawing the dead bones of other , better movies .

Figure 3: An exemplary output of the RTP for a negative review from the movie reviews dataset. Green text indicates the ground-truth annotations.

B.2 INAS Dataset

filling in the gaps : modelling native species richness and invasions using spatially incomplete data . detailed knowledge of patterns of native species richness , an important component of biodiversity, and non - native species invasions is often lacking even though this knowledge is essential to conservation efforts . however , we cannot afford to wait for complete information on the distributionand abundance of native and harmful invasive species . using information from counties well surveyed for plants across the usa , we developed models to fill data gaps in poorly surveyed areasby estimating the density (number of species km (- 2)) of native and non - native plant species . here , we show that native plant species density is non - random , predictable , and is the bestpredictor of non - native plant species density . we found that eastern agricultural sites and coastal areas are among the most invaded in terms of non - native plant species densities , and that thecentral usa appears to have the greatest ratio of non - native to native species . these large - scale models could also be applied to smaller spatial scales or other taxa to set priorities for conservationand invasion mitigation , prevention , and control efforts .

Figure 4: An exemplary output of the RTP for an abstract by Jarnevich et al. (2006), which is included in the INAS dataset. The rationale was created for the *Biotic Resistance Hypothesis* label, with green spans indicating the ground-truth annotations.

influence of insects and fungal pathogens on individual and population parameters of cirsium arvense in its native and introduced ranges . introduced weeds are hypothesized to be invasive in their exoticranges due to release from natural enemies . cirsium arvense (californian , canada , or creeping thistle) is a weed of eurasian origin that was inadvertently introduced to new zealand (nz) ,where it is presently one of the worst invasive weeds . we tested the ' enemy release hypothesis ' (erh) by establishing natural enemy exclusion plots in both the native (europe) and introduced (nz) ranges of c . arvense . we followed the development and fate of individually labelled shoots and recorded recruitment of new shoots into the population over two years . natural enemy exclusion hadminimalimpact on shoot height and relative growth rate in either range . however , natural enemies did have a significant effect on shoot population growth and development in the native range , supportingthe erh . in year one , exclusion of insect herbivores increased mean population growth by 2 . 1 - 3 . 6 shoots m (- 2) , and in year two exclusion of pathogens increased mean population growthby 2 . 7 - 4 . 1 shoots m (- 2) . exclusion of insect herbivores in the native range also increased the probability of shoots developing from the budding to the reproductive growth stage by 4 .0x in the first year , and 13 . 4x in the second year ; but exclusion of pathogens had no effect on shoot development in either year . in accordance with the erh , exclusion of insect herbivores and pathogensdid not benefit shoot development or population growth in the introduced range . in either range , we found no evidence for an additive benefit of dual exclusion of insects and pathogens , andin no case was there an interaction between insect and pathogen exclusion . this study further demonstrates the value of conducting manipulative experiments in the native and introduced ranges of aninvasive plant to elucidate invasion mechanisms .

Figure 5: An exemplary output of the RTP for an abstract by Cripps et al. (2011), which is included in the INAS dataset. The rationale was created for the *Enemy Release Hypothesis* label, with green spans indicating the ground-truth annotations.