A Multilingual Perspective Towards the Evaluation of Attribution Methods in Natural Language Inference

Kerem Zaman
Boğaziçi University
Istanbul, Turkey
kerem.zaman@boun.edu.tr

Yonatan Belinkov
Technion – Israel Institute of Technology
Haifa, Israel
belinkov@technion.ac.il

Abstract

Most evaluations of attribution methods focus on the English language. In this work, we present a multilingual approach for evaluating attribution methods for the Natural Language Inference (NLI) task in terms of plausibility and faithfulness properties. First, we introduce a novel cross-lingual strategy to measure faithfulness based on word alignments, which eliminates the potential downsides of erasure-based evaluations. We then perform a comprehensive evaluation of attribution methods, considering different output mechanisms and aggregation methods. Finally, we augment the XNLI dataset with highlight-based explanations, providing a multilingual NLI dataset with highlights, which may support future exNLP studies. Our results show that attribution methods performing best for plausibility and faithfulness are different.[1]

1 Introduction

The opaqueness of large pre-trained models such as BERT (Devlin et al., 2019) and GPT (Radford and Narasimhan, 2018) motivates the development of explanation methods (Wallace et al., 2020), which aim to attribute importance to particular input features (Ribeiro et al., 2016; Sundararajan et al., 2017; Springenberg et al., 2015; Bach et al., 2015), such as words in a textual input. Two main criteria for evaluating such methods are plausibility and faithfulness (Jacovi and Goldberg, 2020). Plausibility can be defined as the consistency between explanations and human expectations, while faithfulness can be defined as the consistency between explanations and the underlying decision-making process of the model.

Prior evaluations of attributions along these dimensions (Atanasova et al., 2020; DeYoung et al., 2020; Ding and Koehn, 2021) suffer from several limitations. First, they have been limited in (a) the range of considered attribution methods; and (b) the mechanism of calculating the attributions. Second, standard faithfulness evaluations such as erasure-based (DeYoung et al., 2020) suffer from the problem of out-of-distribution examples, where examples presented to the model during attribution are significantly different from those the model has been trained on (Bastings and Filippova, 2020). Third, prior plausibility evaluations are limited to English-only datasets since there is a lack of multilingual datasets with highlighted rationales.

In this work, we aim to fill this gap. Our main contribution is a new framework for evaluating the faithfulness of attribution methods. Inspired by Jacovi and Goldberg (2020)'s criterion for faithful explanations as giving similar explanations for similar inputs, we propose to use cross-lingual sentences (translations) as similar inputs. Given a multilingual model, we argue that faithful attributions should point to words that are aligned in two translations of the same sentence. This approach avoids out-of-distribution inputs by utilizing cross-lingual sentences as *naturally ocurring* input perturbations. We also eliminate the need for carefully crafted and relatively small datasets since our method requires only a multilingual parallel corpus.

We focus on Natural Language Inference (NLI) as a case study, since it is a central task that has been widely used as a test bed for attribution methods (Atanasova et al., 2020; DeYoung et al., 2020; Jain and Wallace, 2019; Kim et al., 2020; Wiegreffe and Marasović, 2021; Prasad et al., 2021). We compare eight attribution methods, including different mechanisms of computation varying the output and the aggregation of input feature importance scores.

First, we experiment with the cross-lingual XNLI dataset (Conneau et al., 2018) and multilingual BERT (Devlin et al., 2019), and discover large differences in the faithfulness of different attribution methods.

Second, we find that certain attributions are more

[1] Our code is available in https://github.com/KeremZaman/explaiNLI.

plausible and that the choice of computation mechanism has a large effect in some cases. As far as we know, this is the first comprehensive study investigating the effect of different types of outputs when evaluating attributions.

Informed by our comprehensive evaluation, we augment the multilingual XNLI dataset (Conneau et al., 2018) with highlight-based explanations by extracting highlights for the English part of XNLI and projecting along word alignments to other languages. We perform a plausibility evaluation with the resulting dataset, which we dub e-XNLI, and perform a human evaluation for a subset of the dataset to validate its adequacy.

Finally, when comparing the ranking of attribution methods by plausibility and faithfulness, we find that no single method performs best. Different methods have different pros and cons, and may therefore be useful in different scenarios. In summary, this work provides:

- A novel faithfulness evaluation framework.
- A comprehensive evaluation of attribution methods, which may guide practitioners when applying such methods.
- A dataset containing explanations in multiple languages for the NLI task, which may support future multilingual exNLP studies.

2 Background

2.1 Properties for Evaluating Attributions

Many properties have been defined to evaluate explanations with respect to different aspects. Plausibility and faithfulness (Jacovi and Goldberg, 2020), sufficiency (DeYoung et al., 2020), stability and consistency (Robnik-Sikonja and Bohanec, 2018), and confidence indication (Atanasova et al., 2020) are examples of such properties. As two prominent ones, we focus on faithfulness and plausibility.

2.1.1 Faithfulness

Faithfulness is the measure of how much an interpretation overlaps with the reasoning process of the model. In other words, if the scores given by an attribution method are compatible with the decision process behind the model, that interpretation is considered faithful. Such compatability may be instantiated in different ways. For example, Ding and Koehn (2021) measure faithfulness through model consistency and input consistency. They measure model consistency by comparing attribution scores of two different models, where one of them is the distilled version of the other. For input consistency, they compare the attribution scores of perturbed input pairs. Perturbing inputs or erasing some parts from input is a widely-used technique for faithfulness evaluation (Arras et al., 2017; Serrano and Smith, 2019; DeYoung et al., 2020; Ding and Koehn, 2021; Atanasova et al., 2020). The basic idea behind these methods is to observe the effect of changing or removing parts of inputs on model output. For instance, if removing words with high attribution scores changes the model output, then the explanation is faithful. For these methods, the change in prediction score is usually assumed to be caused by deletion of the significant parts from the input. However, the main reason might be the out-of-distribution (OOD) inputs created by the perturbations (Bastings and Filippova, 2020). The dependence on perturbations that result in OOD inputs is the main drawback of common faithfulness evaluation methods. In Section 3.1.1 we propose a new evaluation that overcomes this drawback.

2.1.2 Plausibility

Plausibility is a measure of how much an explanation overlaps with human reasoning (Ding and Koehn, 2021). In particular, if an attribution method tends to give higher scores to the part of the inputs that affect the decision according to humans, then it is plausible. In general, human-annotated highlights (parts of the input) are used for plausibility evaluation (Wiegreffe and Marasović, 2021), which we also follow in this work. However, some recent studies use lexical agreement (Ding and Koehn, 2021), human fixation patterns based on eye-tracking measurements (Hollenstein and Beinborn, 2021), and machine translation quality estimation (Fomicheva et al., 2021).

2.2 Overview of Attribution Methods

In this work, we focus on the evaluation of local post-hoc methods, which provide explanations to the output of a model for a particular input by applying additional operations to the model's prediction (Danilevsky et al., 2020). Local post-hoc methods can be grouped into three categories: methods based on gradients, perturbations, or simplification (Atanasova et al., 2020). In gradient-based methods, the gradient of the model's output with respect to the input is used in various ways for calculating attribution scores on the input. Perturbation-based methods calculate attribution scores according to the change in the model's output after perturbing the input in different ways. Simplication-based methods simplify the model to assign attributions. For instance, LIME (Ribeiro et al., 2016) trains a simpler surrogate model covering the local neighborhood of the given input.

The attribution methods we evaluate are as follows: InputXGradient (Shrikumar et al., 2017), Saliency (Simonyan et al., 2014), GuidedBackprop (Springenberg et al., 2015), and IntegratedGradients (Sundararajan et al., 2017) as gradient-based methods; Occlusion (Zeiler and Fergus, 2014) and Shapley Value Sampling (Ribeiro et al., 2016) as perturbation-based; LIME (Ribeiro et al., 2016) as simplification-based; and Layer Activation (Karpathy et al., 2015). We provide details about these methods in Appendix B.

2.3 Output Mechanisms and Aggregation Methods

Most previous studies compute attributions when the output is the top predicted class. We also compare with the case when the output is the loss value calculated with respect to the gold label. More formally, let $f(\mathbf{x}^{(i)})$ denote the output of a classification layer, where $x^{(i)}$ is i-th instance of the dataset. Then, for the common cross-entropy loss, the loss output can be expressed as $y^{(i)}log(f(\mathbf{x}^{(i)}))$ and the top predicted class can be expressed as $\max f(\mathbf{x}^{(i)})$. Furthermore, some attribution methods, such as InputxGradient and Saliency, return importance scores for each dimension of each input word embedding, which need to be aggregated to obtain a single score for each word. While prior studies use different aggregation operations, namely mean and L_2, we examine their effect exhaustively.

Denote the importance score for the k-th dimension of the j-th word embedding of $\mathbf{x}^{(i)}$ as $u^{(i)}_{jk}$. Then we obtain an attribution score per word, $\omega^{(i)}_{\mathbf{x}_j}$, using mean aggregation as follows:

$$\omega^{(i)}_{\mathbf{x}_j} = \frac{1}{N}\sum_{k=0}^{d} u^{(i)}_{jk} \tag{1}$$

where N is the number of words in the given sequence and d is the number of dimensions for the embedding. Similarly, we define the attribution score per word using L_2 aggregation as follows:

$$\omega^{(i)}_{\mathbf{x}_j} = \sqrt{\sum_{k=0}^{d}(u^{(i)}_{jk})^2} \tag{2}$$

3 Methods

3.1 Faithfulness Evaluation

3.1.1 Crosslingual Faithfulness Evaluation

For faithfulness evaluation, erasure-based methods examine the drop in prediction scores by removing the important tokens from the input (Section 2.1.1). On the other hand, the drop in the prediction scores may be the result of the altered, out-of-distribution inputs (Bastings and Filippova, 2020). To overcome this problem, we design a new strategy to evaluate faithfulness by relying on cross-lingual models and datasets. Before diving into details, it is useful to remind Corrolary 2 from Jacovi and Goldberg (2020).

Corrolary 2 *An interpretation system is unfaithful if it provides different interpretations for similar inputs and outputs.*

The main intuition behind our method is to use translation pairs to provide similar inputs to a single model. In particular, we assume a multilingual model that can accept inputs from different languages, such as multilingual BERT (mBERT; Devlin et al. 2019). Then, we can look at the attribution scores of matching parts (words or phrases) of the similar inputs.

This idea consists of several steps. First, we construct translation pairs of which source and target are English and another language, respectively. Second, we calculate attribution scores for instances in English and other languages. Third, the attribution scores are aligned between source and target through word alignments. Finally, attribution scores calculated for English instances are compared with the ones for corresponding words in other languages by calculating the average Spearman correlation between aligned attribution scores. By looking at the correlation between corresponding parts of the inputs, we measure how consistent the model is for similar inputs. Figure 1 illustrates the cross-lingual faithfulness evaluation procedure.

More formally, let $\mathbf{x}^{(i)}_c = \langle x^{(i)}_{c,1}, x^{(i)}_{c,2}, \ldots, x^{(i)}_{c,n}\rangle$ denote the i-th instance of the dataset for language c (out of C languages), where $x^{(i)}_{c,j}$ stands for j-th word of the instance. Let $A = \{(x^{(i)}_{en,k}, x^{(i)}_{c,j}) : x^{(i)}_{en,k} \in \mathbf{x}^{(i)}_{en}, x^{(i)}_{c,j} \in \mathbf{x}^{(i)}_c\}$ be set of words from $\mathbf{x}^{(i)}_c$ that are aligned with words in the corresponding English sentence, $\mathbf{x}^{(i)}_{en}$.[2] Denote by $\omega^{(i)}_{x_{c,j}}$ the

[2] We choose English as the reference language since our

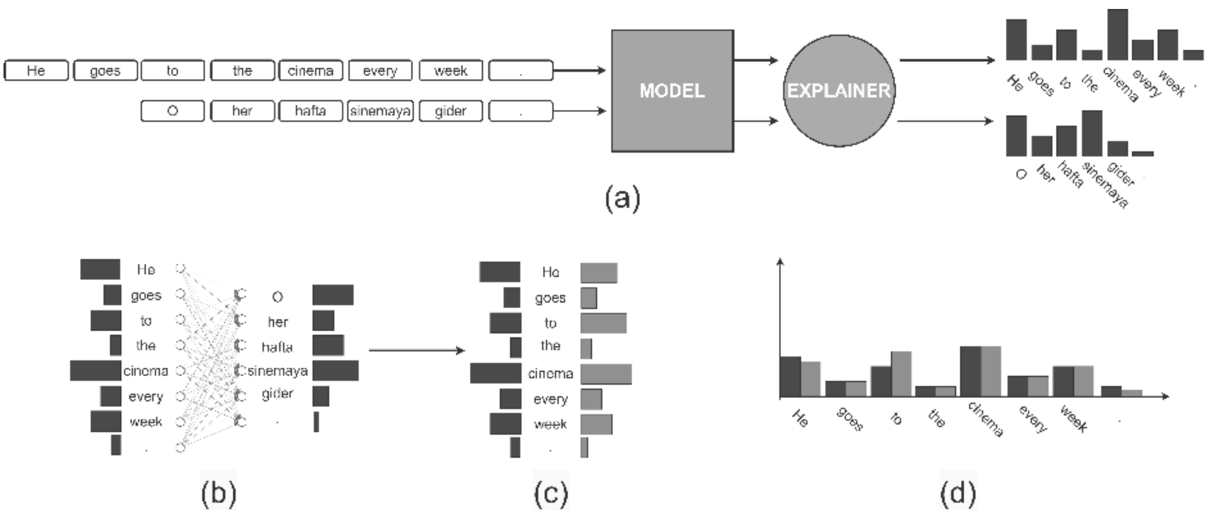

Figure 1: Illustration of cross-lingual faithfulness evaluation. (a) For any en–XX sentence pair (in this example, English–Turkish), we pass each item of the pair through the cross-lingual model and attribution method, to get attribution scores. (b) We extract word alignments by using awesome-align and (c) align scores for the words in Turkish with the ones in the English language by summing the scores of corresponding Turkish words for each English word. (d) Finally, we get two different distributions for the English sentence: the calculated attribution scores and the aligned attribution scores. We compare them to evaluate faithfulness.

attribution score for word $x_{c,j}^{(i)}$ and let $\omega_{\mathbf{x}_c}^{(i)} = \langle \omega_{x_{c,1}}^{(i)}, \omega_{x_{c,2}}^{(i)}, \ldots, \omega_{x_{c,n}}^{(i)} \rangle$. In order to align attribution scores for instances from another languages with the English ones, we define the aligned attribution score for each word in the reference language as the sum of the attribution scores of the corresponding words in the target language:

$$\overline{\omega}_{x_{c,k}}^{(i)} = \sum_{(x_{en,k}^{(i)}, x_{c,j}^{(i)}) \in A} \omega_{x_{c,j}}^{(i)} \tag{3}$$

By aligning scores, we obtain equivalent attribution scores in the target language for each word in the source language. Finally, we define the cross-lingual faithfulness (ρ) of a dataset as the average Spearman correlation between attribution scores for English and aligned attribution scores for all other languages:

$$\rho = \frac{1}{C-1} \frac{1}{M} \sum_{c \neq en} \sum_{i=0}^{M} \rho_{\omega_{\mathbf{x}_{en}}^{(i)}, \overline{\omega}_{\mathbf{x}_c}^{(i)}} \tag{4}$$

The main advantage of this approach is in avoiding the OOD problem: Translation pairs form naturally occurring perturbations that are part of the model's training distribution, unlike the synthetic inputs formed by erasure-based methods. We also reduce the language-specific bias by using translations of the same sentence in different languages.

cross-lingual model performs best on it and since the word aligner we use was originally fine-tuned and evaluated on en–XX language pairs.

3.1.2 Erasure-based Faithfulness Evaluation

To compare our method with erasure-based faithfulness evaluation methods, we report sufficiency and comprehensiveness (DeYoung et al., 2020), which are common metrics for erasure-based faithfulness evaluation, for each attribution method. We stick to their definitions and choices along the experiments.

Let $m(\mathbf{x}^{(i)})_j$ be the model output of the j-th class for the i-th data point and $r^{(i)}$ be the most important tokens to be erased, decided according to attribution scores. Comprehensiveness measures the drop in prediction probability after removing the important tokens (higher values are better):

$$\text{comprehensiveness} = m(\mathbf{x}^{(i)})_j - m(\mathbf{x}^{(i)} \backslash r^{(i)})_j \tag{5}$$

Sufficiency measures the drop when only the important tokens are kept (lower values are better):

$$\text{sufficiency} = m(\mathbf{x}^{(i)})_j - m(r^{(i)})_j \tag{6}$$

$r^{(i)}$ is the top-k_d words according to their attribution scores, where k_d depends on the dataset. However, choosing an appropriate k can be tricky, especially when human rationales are not available to decide an average length. Also, the variable k_d makes scores incomparable across datasets. To solve these issues, they propose Area Over Perturbation Curve (AOPC) metrics for sufficiency and comprehensiveness, where they define bins of tokens to be deleted. They calculate comprehensiveness and sufficiency when top tokens contained by each bin are deleted, then they obtain AOPC measures by averaging the scores for each bin. Here we group the top 1%, 5%, 10%, 20%, 50% tokens into bins in the order of decreasing attribution scores.

3.2 Plausibility Evaluation

To evaluate the plausibility of attribution methods, we measure agreement with human rationales, following Atanasova et al. (2020). This evaluation measures how much the attribution scores overlap with human annotations by calculating Mean Average Precision (MAP) across a dataset. For each instance in the dataset, Average Precision (AP) is calculated by comparing attribution scores $\boldsymbol{\omega}^{(i)}$ with gold rationales, $\mathbf{w}^{(i)}$, where $\boldsymbol{\omega}^{(i)}$ stands for the attribution scores calculated for the dataset instance $\mathbf{x}^{(i)}$ and $\mathbf{w}^{(i)}$ stands for the sequence of binary labels indicating whether the token is annotated as the rationale. For a dataset $X = \{\mathbf{x}^{(i)} | i \in [1, M]\}$, the MAP score is defined as:

$$\mathrm{MAP}(\omega, X) = \frac{1}{M} \sum_{i \in [1,M]} AP(\mathbf{w}^{(i)}, \boldsymbol{\omega}^{(i)}) \quad (7)$$

4 Experiments

4.1 Faithfulness Experiments

Experimental setup We use the XNLI dataset (Conneau et al., 2018) to construct translation pairs where source and target are English and other languages, respectively. We use awesome-align (Dou and Neubig, 2021) to align attribution scores for the corresponding words in translation pairs.[3] As a cross-lingual model, we fine-tune mBERT on the multiNLI dataset (Williams et al., 2018). For cross-lingual faithfulness evaluation, we only use the top-5 languages from XNLI where our fine-tuned mBERT performs best in zero-shot prediction. The cross-lingual performance of our model on all XNLI languages appears in Appendix A.

4.1.1 Cross-lingual Faithfulness Experiments

Table 1 shows cross-lingual faithfulness results for each attribution method, when computing attributions with regard to top prediction or loss, and when aggregating input scores with L_2 or mean aggregation. The results exhibit a large variation, indicating that our cross-lingual faithfulness evaluation is able to expose differences between attribution methods. Activation with mean aggregation is the most faithful attribution method for both types of attribution calculation. We also observe that gradient-based attribution methods (first 8 rows in Table 1) usually generate more faithful explanations than perturbation-based ones (last two rows), in line with prior work (Atanasova et al., 2020).

[3]We use the model provided by the authors, which was multilingually fine-tuned without consistency optimization, due to its good zero-shot performance.

Method	ρ	
	TP	**Loss**
InputxGradient (μ)	.0547	.0746
InputxGradient (L_2)	.6836	.6851
Saliency (μ)	.6124	.6145
Saliency (L_2)	.6129	.615
GuidedBackProp (μ)	.0034	.0015
GuidedBackProp (L_2)	.6129	.615
IntegratedGrads (μ)	.1703	.2546
IntegratedGrads (L_2)	.5884	.5226
Activation (μ)	**.6882**	**.6882**
Activation (L_2)	.6878	.6878
LIME	.0733	.0943
Occlusion	.1514	.306
Shapley	.3418	.4454

Table 1: Cross-lingual faithfulness results: Average correlations measured for different attribution methods on the XNLI dataset. Attribution calculations are performed with respect to the top prediction (TP) class and the loss. Activation with mean aggregation (μ) is the best performing method in both cases.

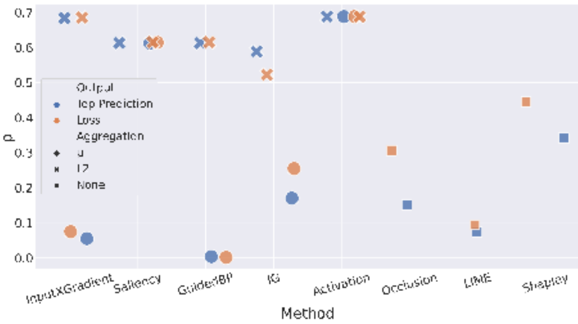

Figure 2: Comparison of cross-lingual faithfulness along output and aggregation dimensions. L_2 mostly outperforms mean (μ) aggregation and calculations with respect to the loss are the same or slightly better than ones with respect to the top predicted class.

Figure 2 shows the effect of the aggregation methods and output mechanisms on cross-lingual faithfulness. For all cases, L_2 aggregation outperforms the mean aggregation by large margins except Saliency and Activation. While the score for mean aggregation is very close to L_2 aggregation for Saliency, it is slightly better than L_2 aggregation for Activation. Since Saliency returns the absolute value, it does not contradict the general trend for the effect of L_2 aggregation on gradient-based attribution methods as in plausibility evaluation. Considering output mechanisms, calculating attribution scores with respect to loss is the same or slightly better than the ones with respect to the top predicted class in almost all cases. For Integrated Gradients with L_2 aggregation and GuidedBackprop with mean aggregation, calculating attribution scores with respect to the top predicted class performs better.

Recall that our cross-lingual faithfulness measure averages correlations across languages (Eq. 4). To analyze the effect of languages, Table 2 shows correlations per language when averaged across all combinations of methods, outputs and aggregations. The results show little variation across languages, although languages with better NLI performance tend to yield more faithful explanations. Detailed results per language and attribution method are available in Appendix C.

	de	es	fr	vi	zh
ρ	.43	.46	.45	.40	.37
F1	.72	.74	.74	.70	.70

Table 2: Cross-lingual faithfulness results (ρ) per language averaged across all attribution methods on the XNLI dataset, and NLI F1 scores for comparison.

4.1.2 Erasure-based Faithfulness Experiments

Table 3 shows the results of erasure-based faithfulness evaluation (comprehensiveness and sufficiency), for each attribution method. According to the results, InputxGradient with L_2 aggregation is the most faithful attribution method in terms of comprehensiveness when the output is the top prediction class; Saliency and GuidedBackpropagation methods with L_2 aggregation are the most faitful ones in terms of comprehensiveness when the output is the loss. For sufficiency, Activation seems to be the most faithful method for both cases. Interestingly, most of the results are quite similar and differences between methods are not as large as in the cross-lingual faithfulness evaluation.

Figure 3 shows the effect of aggregation method and output mechanism on comprehensiveness. For all attribution methods, L_2 outperforms mean aggregation except for Saliency with top prediction

Method	comp. ↑		suff. ↓	
	TP	Loss	TP	Loss
InputxGradient (μ)	.2849	.2964	.2666	.2423
InputxGradient (L_2)	**.3222**	.3148	.2358	.2613
Saliency (μ)	.3139	.3184	.2259	.2319
Saliency (L_2)	.3098	**.3206**	.2383	.2377
GuidedBackprop (μ)	.2737	.2052	.2817	.2862
GuidedBackprop (L_2)	.3098	**.3206**	.2383	.2377
IntegratedGrads (μ)	.2128	.2586	.2881	.2134
IntegratedGrads (L_2)	.3021	.291	.2907	.2872
Activation (μ)	.2402	.2402	**.179**	**.179**
Activation (L_2)	.3065	.3065	.333	.333
LIME	.2449	.2493	.241	.2261
Occlusion	.2986	.307	.2891	.2382
Shapley	.3045	.3129	.2756	.2219

Table 3: Erasure-based faithfulness results: Average AOPC comprehensiveness and sufficiency scores for different attribution methods on the English split of XNLI. Attribution calculations are performed with respect to the top predicted class (TP) and the loss. For comprehensiveness, InputxGradient with L_2 aggregation performs best when attributions are calculated with respect to top prediction, while Saliency and Guided Backpropagation with L_2 aggregation perform best when calculating with respect to the loss. For sufficiency, Activation with mean aggregation performs best in both cases.

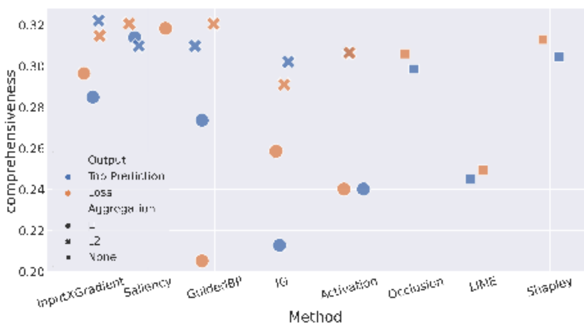

Figure 3: Comparison of comprehensiveness results along output and aggregation dimensions. L_2 outperforms mean aggregation and calculations with respect to the loss slightly outperform calculations with respect to the top prediction class for most attribution methods.

class as output. In almost all cases, calculating attribution scores with respect to loss is as good as or slightly better than calculating with respect to the top predicted class. For InputxGradient with L_2 aggregation and Guided Backprop with mean aggregation, calculating attributions with respect to the top predicted class performs better.

Figure 4 shows the effect of the aggregation method and output mechanism on sufficiency. Unlike comprehensiveness, mean aggregation outperforms L_2 aggregation for most attribution methods

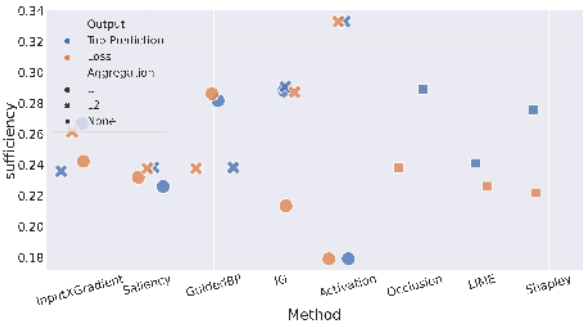

Figure 4: Comparison of sufficiency results along output and aggregation dimensions. Mean (μ) outperforms L_2 aggregation and calculations with respect to loss slightly outperform or are the same as those calculated with respect to top prediction for most attribution methods.

except for InputXGradient with top prediction as output and both GuidedBackprop methods. Calculating attribution scores with respect to loss is the same or slightly better than the ones with respect to the top predicted class except GuidedBackprop with mean aggregation, InputxGradient with L_2 aggregation and Saliency with mean aggregation.

4.1.3 Cross-lingual vs. Erasure-based Faithfulness

The results of cross-lingual faithfulness and erasure-based metrics (comprehensiveness and sufficiency) differ in two main aspects:

- Perturbation-based methods exhibit more faithful explanations when evaluated by erasure-based metrics than when evaluated by cross-lingual faithfulness. We interpret this pattern as a result of the OOD issue caused by erasure-based evaluation, which unjustifiably favors perturbation-based attributions. The relative improvement for perturbation-based methods can be attributed to noise due to the OOD perturbations used for calculating comprehensiveness and sufficiency.
- Erasure-based faithfulness metrics are unable to properly distinguish between different attribution methods, since the differences are dwarfed by the noise introduced by the OOD perturbations. The standard deviation of faithfulness scores across all attribution methods is 0.26 for cross-lingual faithfulness, but only 0.03 and 0.04 for comprehensiveness and sufficiency, respectively.

4.2 Plausibility Experiments

Experimental Setup We use the e-SNLI dataset (Camburu et al., 2018) to obtain human annotations.

Method	**MAP**	
	TP	**Loss**
InputxGradient (μ)	.385	.392
InputxGradient (L_2)	.636	.643
Saliency (μ)	.645	**.655**
Saliency (L_2)	**.646**	**.655**
GuidedBackProp (μ)	.407	.410
GuidedBackProp (L_2)	**.646**	**.655**
IntegratedGrads (μ)	.470	.339
IntegratedGrads (L_2)	.626	.639
Activation (μ)	.230	.230
Activation (L_2)	.451	.451
LIME	.451	.273
Occlusion	.542	.277
Shapley	.565	.268

Table 4: Plausibility results: MAP scores for different attribution methods on the e-SNLI dataset. Attribution calculations are performed with respect to the top prediction class (TP) and the loss. Saliency with both aggregations and GuidedBackprop with L_2 aggregation are the best performing methods in both cases.

As the classifier, we use a BERT-base model fine-tuned on the SNLI dataset (Bowman et al., 2015), provided by TextAttack (Morris et al., 2020).

Results According to the results (Table 4), Saliency and GuidedBackprop with L_2 aggregation are the most plausible attribution methods for both types of attribution calculation, and Saliency with mean aggregation is one of the most plausible methods when attributing with respect to the loss. Similar to cross-lingual faithfulness results, we observe that gradient-based attribution methods usually generate more plausible explanations than perturbation-based ones, as in prior work (Atanasova et al., 2020).

Figure 5 shows the effect of aggregation method and output mechanism on plausibility. In all cases, L_2 outperforms mean aggregation by large margins except for Saliency, where the score for mean aggregation is very close to L2 aggregation. When we consider that Saliency returns the absolute value, which is analogous to L_1 aggregation, the exception in the results makes sense. In almost all cases, calculating attribution scores with respect to loss is the same or slightly better than calculating with respect to the top predicted class. For Integrated Gradients with mean aggregation, Occlusion, and LIME, calculating attribution scores with respect

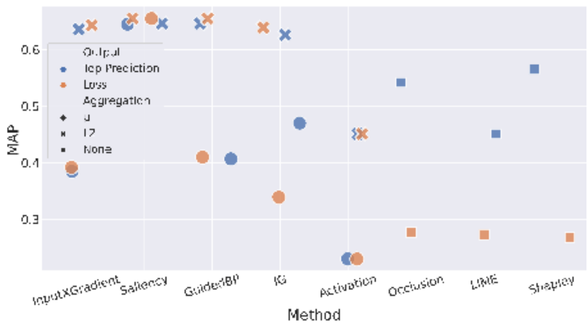

Figure 5: Comparison of plausibility results along output and aggregation dimensions. L_2 outperforms mean aggregation for all attribution methods and calculating attributions with respect to loss is the same or slightly better than with respect to the top predicted class.

Lang	MAP	Lang	MAP	Lang	MAP
ar	0.663	es	0.766	th	0.932
bg	0.701	fr	0.739	tr	0.665
de	0.732	hi	0.604	ur	0.575
el	0.696	ru	0.686	vi	0.572
en	1.0	sw	0.58	zh	0.543

Table 5: Plausibility results: MAP scores measured on the newly introduced e-XNLI dataset (using Saliency with loss as output and L_2 aggregation).

to the loss performs better.

e-XNLI dataset Since prior studies for plausibility evaluation are limited to English-only datasets for NLI task, we augment the XNLI dataset (Conneau et al., 2018) with highlight-based explanations by utilizing the best attribution method for plausibility according to our results. We extract rationales from the English split of the XNLI dataset and align them to other languages using awesome-align. For extracting rationales, we binarize the continuous attribution scores with respect to the threshold that gives the best F1 score on the e-SNLI dataset. We choose Saliency with L_2 aggregation and loss as output for calculating attribution scores since it is one of the two most plausible methods.

To validate the automatically generated highlights, we follow two approaches. First, we measure the plausibility of the same attribution method used to extract rationales for those languages. This approach investigates whether the aligned rationales are able to follow the same reasoning paths for each language. As Table 5 shows, the automatically aligned highlights in e-XNLI are similarly plausible explanations for most languages.

Language	Precision	Recall	F1
ar	.64	.73	.68
en	.79	.78	.79
ru	.93	.78	.85
tr	.77	.71	.74

Table 6: Human evaluation for a sample of e-XNLI: Precision, recall and F1 scores for four languages.

Second, we perform a human evaluation on a subset of the created dataset. For four XNLI languages, we sample 10 examples per label (30 total) and request annotators to evaluate the correctness of highlight by following the same procedure carried out in e-SNLI (Camburu et al., 2018). Then, we measure precision, recall, and F1 scores between automatically generated highlights and those manually edited by human annotators. As Table 6 shows, automatically generated highlights mostly agree with human reasoning.

We make the e-XNLI dataset publicly available under MIT license to facilitate research on explainable NLP in a multilingual setting. [4]

5 Conclusion

We introduce a novel cross-lingual strategy to evaluate the faithfulness of attribution methods, which eliminates the out-of-distribution input problem of common erasure-based faithfulness evaluations. Then, we perform a comprehensive comparison of different attribution methods having different characteristics in terms of plausibility and faithfulness. The experiments show that there is no one-size-fits-all solution for local post-hoc explanations. Our results highlight that practitioners should choose an attribution method with proper output mechanism and aggregation method according to the property of explanation in question:

- For most attribution methods, L_2 aggregation and attribution calculation with respect to loss provide more faithful and plausible explanations.
- Erasure-based faithfulness metrics cannot properly differentiate different attribution methods.
- Gradient-based attribution methods usually generate more plausible and faithful explanations than perturbation-based methods.
- One should choose Guided Backpropagation

[4] https://github.com/KeremZaman/explaiNLI

with L_2 and Saliency with both aggregation methods and calculate scores with respect to the loss to obtain the most plausible explanations.

- One should choose Activation with L_2 regardless of output mechanism to obtain the most faithful explanations.

Finally, we present e-XNLI, a multilingual dataset with automatically generated higlight explanations, to support future multilingual exNLP studies.

Acknowledgements

We would like to thank Oleg Serikov for evaluating automatically extracted highlights in the Russian subset of our e-XNLI dataset and Ramazan Pala for reviews while drafting this paper. This research was partly supported by the ISRAEL SCIENCE FOUNDATION (grant No. 448/20) and by an Azrieli Foundation Early Career Faculty Fellowship. YB is supported by the Viterbi Fellowship in the Center for Computer Engineering at the Technion.

References

Leila Arras, F. Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. "what is relevant in a text document?": An interpretable machine learning approach. *PLoS ONE*, 12.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

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.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *NeurIPS*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

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.

Shuoyang Ding and Philipp Koehn. 2021. Evaluating saliency methods for neural language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5034–5052, Online. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

M. Fomicheva, Lucia Specia, and Nikolaos Aletras. 2021. Translation error detection as rationale extraction. *ArXiv*, abs/2108.12197.

Nora Hollenstein and Lisa Beinborn. 2021. Relative importance in sentence processing. In *Proceedings of the 59th Annual Meeting of the Association* *for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 141–150, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 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*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *NAACL*.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *ArXiv*, abs/1506.02078.

Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. 2020. Interpretation of NLP models through input marginalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3154–3167, Online. Association for Computational Linguistics.

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.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. 2021. To what extent do human explanations of model behavior align with actual model behavior? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 1–14, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

M. Robnik-Sikonja and Marko Bohanec. 2018. Perturbation-based explanations of prediction models. In *Human and Machine Learning*.

Hassan Sajjad, Narine Kokhlikyan, Fahim Dalvi, and Nadir Durrani. 2021. Fine-grained interpretation and causation analysis in deep NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, Online.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

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. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. 2015. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*.

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.

Erik Štrumbelj and Igor Kononenko. 2010. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.*, 11:1–18.

Eric Wallace, Matthew Thomas Gardner, and Sameer Singh. 2020. Interpreting predictions of nlp models. In *EMNLP*.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. In *Proceedings of NeurIPS*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *ECCV*.

A Cross-lingual performance of mBERT classifier

Table 7 shows the results of the mBERT model fine-tuned on multiNLI for each language in the XNLI dataset.

Language	F1
ar	0.6534
bg	0.6815
de	0.7169
el	0.6655
en	**0.8153**
es	0.7426
fr	0.7426
hi	0.6169
ru	0.6767
sw	0.5165
th	0.5289
tr	0.6486
ur	0.5819
vi	0.6992
zh	0.7016

Table 7: F1 scores of the mBERT model fine-tuned on multiNLI for each XNLI language.

B Attribution Methods

In this work, we focus on a wide range of attribution methods by investigating different combinations of output mechanisms and aggregation methods. We consider two different output options while calculating importance scores per word: (a) top predicted class; (b) loss value calculated when the ground truth label is given. In the following, we refer to the output as f_{tp} when it is the top predicted class and $f_{\mathcal{L}}$ when it is the loss. While some methods inherently return a single score per word, some of them return importance scores for each dimension of the corresponding word vector. Since we want to obtain a single score per word, those scores are need to be aggregated. We investigate L_2 and mean aggregations separately.

Implementation Details We build our framework upon the Captum library (Kokhlikyan et al., 2020) to use existing implementations of many attribution methods. We use the HuggingFace transformers (Wolf et al., 2020) and datasets (Lhoest et al., 2021) libraries to access pretrained models and datasets. Also, we rely upon Scikit-learn (Pedregosa et al., 2011) for evaluation scores such as Average Precision (AP) and Spearman Correlation.

B.1 Saliency

Saliency (Simonyan et al., 2014) calculates attribution scores by calculating the absolute value of the gradients with respect to inputs. More formally, let u_j be the embedding for word x_j of $\mathbf{x}^{(i)}$, the i'th instance of any dataset. Then the attribution score per each dimension of the embedding is defined as

$$|\nabla_{u_{jk}} f(\mathbf{x}^{(i)})| \tag{8}$$

We obtain an attribution score per word, $\omega_{x_j}^{(i)}$, by aggregating scores across each word embedding. Using mean aggregation, it is defined as follows:

$$\omega_{x_j}^{(i)} = \frac{1}{N} \sum_{k=0}^{d} |\nabla_{u_{jk}} f(\mathbf{x}^{(i)})| \tag{9}$$

where d is the number of dimensions for the word embedding and N is number of words in the sequence. Similarly, using L_2 aggregation, we obtain

$$\omega_{x_j}^{(i)} = \sqrt{\sum_{k=0}^{d} |\nabla_{u_{jk}} f(\mathbf{x}^{(i)})|^2} \tag{10}$$

B.2 InputxGradient

InputxGradient (Shrikumar et al., 2017) calculates attribution scores by multiplying the input with the gradients with respect to the input. More formally, the attribution score per each dimension is defined as

$$\nabla_{u_{jk}} f(\mathbf{x}^{(i)}) u_{jk} \tag{11}$$

We obtain attribution scores per word in the same way as Saliency using mean/L_2 aggregations.

B.3 Guided Backpropagation

Guided Backpropagation (Springenberg et al., 2015) produces attribution scores by calculating gradients with respect to the input. Different from other methods, it overrides the gradient of the ReLU activation so that only positive gradients pass through. We obtain attribution scores per word using L_2 and mean aggregations as in the previously described methods.

B.4 Integrated Gradients

Integrated Gradients (Sundararajan et al., 2017) produces attribution scores by summing gradients along each dimension from some baseline input to a given input. The attribution score per each dimension is defined as

$$u_{jk}^{(i)} - \overline{u}_{jk}^{(i)} \times \sum_{l=1}^{m} \frac{\partial f(\overline{u}_{jk}^{(i)} + \frac{l}{m} \times (u_{jk}^{(i)} - \overline{u}_{jk}^{(i)}))}{\partial u_{jk}^{(i)}} \times \frac{1}{m} \tag{12}$$

where m is the number of steps for a Riemannian approximation of the path integral and $\overline{u}_j^{(i)}$ is the baseline input. We use the word embedding of the [PAD] token as the baseline input for each word except for [SEP] and [CLS] tokens (Sajjad et al., 2021). We obtain attribution scores per word using L_2 and mean aggregations as in the previous methods.

Higher values of m would produce a better approximation, but also make attribution calculation computationally expensive. We need to find a sweet spot between approximation and computational resources. For plausibility experiments, we select m according to validation performance based on MAP scores. Among $\{50, 75, 100, 125\}$, we choose $m = 50$ for calculations with respect to the loss, $m = 75$ for mean aggregation, and $n = 100$ for L_2 aggregation on calculations with respect to top prediction. For cross-lingual faithfulness experiments, we select m according to the evaluation on the validation set based on the Spearman correlation coefficient values. Among {50,75,100}, we choose $m = 100$ for all calculations except for the one with respect to loss with mean aggregation, for which we choose $m = 75$. For erasure-based faithfulness experiments, we use the same values of m for the sake of a fair comparison.

B.5 LIME

LIME (Ribeiro et al., 2016) produces attribution scores by training a surrogate linear model using the points around the input created by perturbing the input and output of perturbations from the original model. A random subset of the input is replaced by a baseline value to create perturbations. We use the word embedding of the [PAD] token as the baseline value (as in Integrated Gradients). Since we create the perturbations by replacing whole word vectors, we obtain a single score per word, which eliminates the need for aggregation. We use 50 samples for training the surrogate model as the default value for the LIME implementation in Captum.

B.6 Occlusion

Occlusion (Zeiler and Fergus, 2014) produces attribution scores by calculating differences in the output after replacing the input with baseline values over a sliding window. We select the shape of the sliding window so that it occludes only the embedding of one word at a time, and we use the word embedding of the [PAD] token as a baseline value (as in Integrated Gradients and LIME). Since we create the perturbations by replacing whole word vectors, we obtain a single score per word.

B.7 Shapley Value Sampling

In Shapley Value Sampling (Štrumbelj and Kononenko, 2010), we take a random permutation of input, which is word embeddings of input sequence in our case, and add them one by one to a given baseline, embedding vector for [PAD] token in our case, to produce attribution score by calculating the difference in the output. The scores are averaged across several samples. We choose the feature group so that one score corresponds to a single word, which eliminates the need for aggregation. We take 25 samples for calculating attributions as the default value for Shapley Value Sampling implementation in Captum.

B.8 Activation

Layer Activation (Karpathy et al., 2015) produces attribution scores by getting the activations in the output of the specified layer. We select the embedding layer for this purpose, which yields an attribution score per each dimension of the embedding equal to u_{jk}. Then, we obtain attribution scores per word using L_2 and mean aggregations as in other methods.

C Cross-lingual Faithfulness Results per Language

Our cross-lingual faithfulness evaluation averages correlations across languages. For completeness, we provide in Tables 8–12 the results of cross-lingual faithfulness evaluation per language.

D Human Evaluation for e-XNLI

A subset of our dataset is evaluated by NLP researchers—the authors and a colleague of one of the authors—from Turkey, Israel, and Russia.

The annotators followed the e-SNLI (Camburu et al., 2018) guidelines for evaluating automatically extracted highligh-based explanations.

E Limitations and Potential Risks

In this work, we examine a wide range of attribution methods along output and aggregation dimensions. However, our experiments are only limited to BERT (Devlin et al., 2019) architecture. The

Method	ρ	
	TP	**Loss**
InputxGradient (μ)	.0524	.0705
InputxGradient (L_2)	**.706**	**.708**
Saliency (μ)	.6177	.6202
Saliency (L_2)	.6186	.6207
GuidedBackProp (μ)	.0034	-0.001
GuidedBackProp (L_2)	.6186	.6207
IntegratedGrads (μ)	.1759	.265
IntegratedGrads (L_2)	.602	.5381
Activation (μ)	.6963	.6963
Activation (L2)	.7011	.7011
LIME	.0759	.0995
Occlusion	.2262	.3156
Shapley	.363	.4658

Table 8: Cross-lingual faithfulness results for the German split of XNLI dataset: Average correlations measured for different attribution methods on the XNLI dataset. Attribution calculations are performed with respect to the top prediction (TP) class and the loss.

multilingual dataset we provide, e-XNLI, consists of automatically extracted highlight-based explanations and should be used with caution for future exNLP studies since we only perform the human evaluation on a small subset of the all dataset. Especially, training self-explanatory models with this dataset can cause undesired outcomes such as poor explanation quality.

F Computational Resources

We mainly use Google Colab for the experiments and Titan RTX in some cases. All experiments for gradient-based attribution methods and Activation take a period of time ranging from 5 minutes to 1 hour, while perturbation-based approaches take several hours. Especially, experiments for Shapley Value Sampling take a few days since its implementation does not use batched operations.

Method	ρ	
	TP	**Loss**
InputxGradient (μ)	.0742	.0933
InputxGradient (L_2)	.7322	.7332
Saliency (μ)	.658	.6591
Saliency (L_2)	.6584	.6595
GuidedBackProp (μ)	.0079	-.0006
GuidedBackProp (L_2)	.6584	.6595
IntegratedGrads (μ)	.1962	.2763
IntegratedGrads (L_2)	.637	.5657
Activation (μ)	**.7341**	**.7341**
Activation (L2)	.7232	.7232
LIME	.0796	.0998
Occlusion	.2612	.3446
Shapley	.3696	.4734

Table 9: Cross-lingual faithfulness results for the French split of XNLI dataset: Average correlations measured for different attribution methods on the XNLI dataset. Attribution calculations are performed with respect to the top prediction (TP) class and the loss.

Method	ρ	
	TP	**Loss**
InputxGradient (μ)	.0756	.1029
InputxGradient (L_2)	.7195	.72
Saliency (μ)	.6595	.6615
Saliency (L_2)	.6598	.6619
GuidedBackProp (μ)	-.0007	.0037
GuidedBackProp (L_2)	.6598	.6619
IntegratedGrads (μ)	.2072	.2965
IntegratedGrads (L_2)	.6238	.5581
Activation (μ)	**.7528**	**.7528**
Activation (L2)	.707	.707
LIME	.0865	.1054
Occlusion	.2739	.3618
Shapley	.3616	.4781

Table 10: Cross-lingual faithfulness results for the Spanish split of XNLI dataset: Average correlations measured for different attribution methods on the XNLI dataset. Attribution calculations are performed with respect to the top prediction (TP) class and the loss.

Method	ρ	
	TP	**Loss**
InputxGradient (μ)	.0441	.0648
InputxGradient (L_2)	.6486	.6503
Saliency (μ)	.5809	.5823
Saliency (L_2)	.5813	.5827
GuidedBackProp (μ)	.0023	.0032
GuidedBackProp (L_2)	.5813	.5827
IntegratedGrads (μ)	.1594	.2473
IntegratedGrads (L_2)	.5597	.4949
Activation (μ)	.6627	.6627
Activation (L2)	**.6748**	**.6748**
LIME	.0627	.085
Occlusion	.1942	.2705
Shapley	.3197	.4235

Table 11: Cross-lingual faithfulness results for the Vietnamese split of XNLI dataset: Average correlations measured for different attribution methods on the XNLI dataset. Attribution calculations are performed with respect to the top prediction (TP) class and the loss.

Method	ρ	
	TP	**Loss**
InputxGradient (μ)	.0273	.0413
InputxGradient (L_2)	.6119	.6139
Saliency (μ)	.5458	.5495
Saliency (L_2)	.5462	.5501
GuidedBackProp (μ)	.004	.0021
GuidedBackProp (L_2)	.5462	.5501
IntegratedGrads (μ)	.1126	.188
IntegratedGrads (L_2)	.5197	.4563
Activation (μ)	.5949	.5949
Activation (L2)	**.6331**	**.6331**
LIME	.0619	.0819
Occlusion	.1615	.2374
Shapley	.2953	.3862

Table 12: Cross-lingual faithfulness results for the Chinese split of XNLI dataset: Average correlations measured for different attribution methods on the XNLI dataset. Attribution calculations are performed with respect to the top prediction (TP) class and the loss.