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

Kerem Zaman*
UNC Chapel Hill
kzaman@cs.unc.edu

Yonatan Belinkov[†]
Technion – Israel Institute of Technology
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 faithfulness and plausibility. First, we introduce a novel cross-lingual strategy to measure faithfulness based on word alignments, which eliminates the drawbacks 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, to 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 like BERT (Devlin et al., 2019) and GPT (Radford and Narasimhan, 2018) motivates developing explanation methods (Wallace et al., 2020), which aim to attribute importance to particular input features (Springenberg et al., 2015; Bach et al., 2015; Ribeiro et al., 2016; Sundararajan et al., 2017), 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 is defined as the consistency between explanations and the model's underlying decision-making process.

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 the range of considered attribution methods and the mechanism of calculating the attributions. Second, standard faithfulness evaluations, such as erasure-based ones (DeYoung et al., 2020), entail running the model on examples outside of the training distribution (Bastings and Filippova, 2020). Third, prior plausibility evaluations are limited to English-only datasets due to the lack of multilingual datasets with highlighted explanations.

In this work, we aim to fill these gaps. 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 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), multilingual BERT (Devlin et al., 2019), and XLM-R (Conneau et al., 2020), and discover large differences in the faithfulness of different attribution methods. Second, we find that certain attributions are more 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 in-

---

vestigating 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 on 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, such as 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). 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, the interpretation is considered faithful. Such compatability may be instantiated in different ways. For instance, Ding and Koehn (2021) measure faithfulness through model consistency and input consistency. For model consistency, they compare attribution scores of a given model and its distilled version. For input consistency, they compare attribution scores of perturbed input pairs.

Perturbing inputs or erasing parts of the 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 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 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 gives higher scores to the part of the inputs that affect the decision according to humans, then it is plausible. Typically, 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. Other post-hoc methods outside of these 3 categories (Kokhlikyan et al., 2020) include Layer Activation (Karpathy et al., 2015), which uses activations of each neuron in the output of a specific layer, and NoiseTunnel (Smilkov et al., 2017; Adebayo et al., 2018).

The attribution methods we evaluate are: 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). Details about these methods appear in Appendix B.

2.3 Output Mechanisms and Aggregation Methods

Most previous studies compute attributions when the output is the top prediction score. 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, the score of the top predicted class can be expressed $\max f(\mathbf{x}^{(i)})$. We also compare with the case when the output is the loss value calculated with respect to the gold label. For the common cross-entropy loss, the loss output can be expressed as $y^{(i)} log(f(\mathbf{x}^{(i)}))$ where $y^{(i)}$ is the gold label. 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 $u_{jk}^{(i)}$. Then we obtain an attribution score per word, $\omega_{\mathbf{x}_j}^{(i)}$, using mean aggregation as follows:

$$\omega_{\mathbf{x}_j}^{(i)} = \frac{1}{d} \sum_{k=0}^{d} u_{jk}^{(i)} \qquad (1)$$

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

$$\omega_{\mathbf{x}_j}^{(i)} = \sqrt{\sum_{k=0}^{d} (u_{jk}^{(i)})^2}. \qquad (2)$$

2.4 Natural Language Inference

Natural Language Inference (NLI) is a well-established Natural Language Understanding (NLU) task where the objective is deciding the relation between given sentence pairs (Consortium et al., 1996; Condoravdi et al., 2003; Bos and Markert, 2005; Dagan et al., 2005; MacCartney and Manning, 2009; Poliak, 2020). When a sentence pair is given, namely a premise and a hypothesis, there are 3 possible outcomes: (i) premise entails hypothesis; (ii) premise and hypothesis contradict; or (iii) they are neutral. This setting makes the task suitable to be modeled as a text classification task.

Although there are many human-annotated NLI datasets, we focus on the MNLI (Williams et al., 2018b), XNLI (Conneau et al., 2018) and e-SNLI (Camburu et al., 2018) datasets. MNLI is a collection of 433K sentence pairs from 10 genres of written and spoken English where pairs are labeled as *entailment*, *contradiction* or *neutral*. This dataset is also part of a general NLU benchmark called GLUE (Wang et al., 2018). XNLI is the cross-lingual extension of the MNLI dataset in which sentence pairs from the validation and test sets of MNLI are translated into 15 languages. the e-SNLI dataset is the enhanced version of SNLI (Bowman et al., 2015a), an English-only NLI dataset having the same format as MNLI, with human-annotated explanations in the form of highlights.

3 Faithfulness

3.1 Evaluation Methods

3.1.1 Crosslingual Faithfulness Evaluation

In faithfulness evaluation, erasure-based methods examine the drop in prediction scores by removing the important tokens from the input (Section 2.1.1). However, 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, let us recall 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 key 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
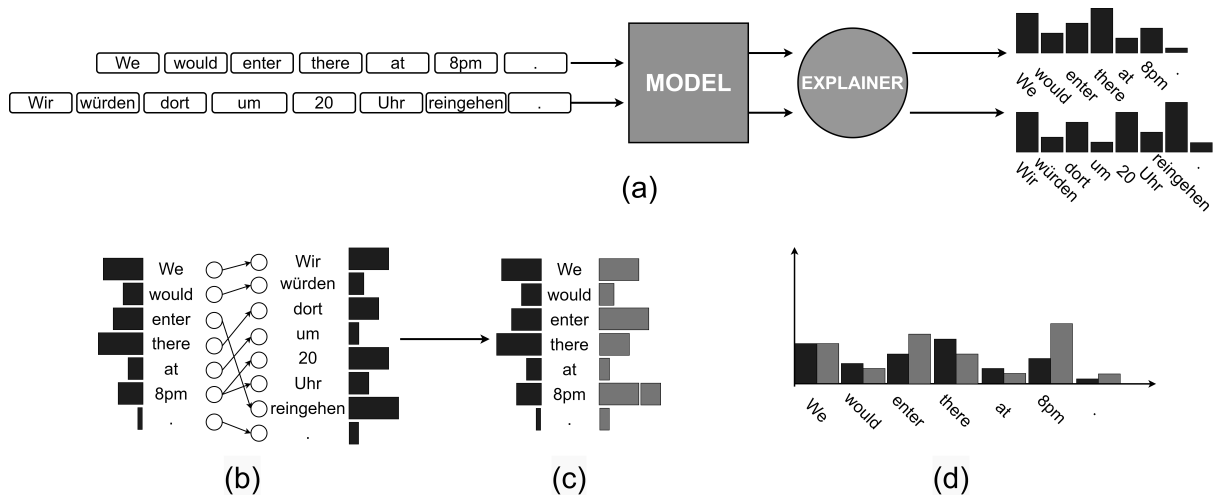
Figure 1: Illustration of cross-lingual faithfulness evaluation. (a) For any en–XX sentence pair (in this example, English–German), we pass each item of the pair through the cross-lingual model and attribution method, to get attribution scores. (b) We extract word alignments by using `awesome-align` and (c) align scores for the words in German with the ones in the English language by summing the scores of corresponding German 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.

languages, such as multilingual BERT (mBERT; Devlin et al. 2019). Then, we can examine the attribution scores of matching parts (words or phrases) of the similar inputs.[2]

This idea consists of several steps. First, construct multiway translation pairs of which source and targets are English and another languages, respectively. Second, calculate attribution scores for instances in English and other languages. Third, align the attribution scores between source and target through word alignments. Finally, correlate attribution scores computed for English instances with the ones for corresponding words in other languages. 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 $\mathbf{x}_c^{(i)} = \langle x_{c,1}^{(i)}, x_{c,2}^{(i)}, \ldots, x_{c,n}^{(i)} \rangle$ denote the i-th instance of the dataset for language c (out of C languages), where $x_{c,j}^{(i)}$ stands for the j-th word of the instance. Let $A = \{(x_{en,k}^{(i)}, x_{c,j}^{(i)}) : x_{en,k}^{(i)} \in \mathbf{x}_{en}^{(i)}, x_{c,j}^{(i)} \in \mathbf{x}_c^{(i)}\}$ be the set of words from $\mathbf{x}_c^{(i)}$ that are aligned with words in the corresponding English sentence, $\mathbf{x}_{en}^{(i)}$.[3] Denote by

$\omega_{x_{c,j}}^{(i)}$ the 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 language 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:

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

By aligning scores, we obtain equivalent attribution scores in the target language for each word in the source language. For the example in Figure 1, we have $\overline{\omega}_{8pm}^{(i)} = \omega_{20}^{(i)} + \omega_{Uhr}^{(i)}$, because $\{(8pm, 20), (8pm, Uhr)\} \subset A$.

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:

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

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

---

[2]We investigate the similarity of translation pairs via their multilingual representations in Appendix C, finding that translation pairs do form similar inputs for a multilingual model.

[3]We use English as the reference language since our 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.

methods. We also reduce language-specific bias by using translations of the same sentence in different languages. Furthermore, our approach provides a grayscale notion of faithfulness, as advocated by Jacovi and Goldberg (2020).

3.1.2 Erasure-based Faithfulness Evaluation

We compare our method with erasure-based faithfulness evaluation metrics, namely sufficiency and comprehensiveness (DeYoung et al., 2020). We stick to DeYoung et al.'s 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):

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

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

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

$r^{(i)}$ 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, DeYoung et al. propose Area Over Perturbation Curve (AOPC) metrics for sufficiency and comprehensiveness, based on bins of tokens to be deleted. They calculate comprehensiveness and sufficiency when deleting the top tokens in each bin, and obtain AOPC metrics 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 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.[4]

---

[4] We use the model provided by the authors, which was multilingually fine-tuned without consistency optimization, due to its good zero-shot performance. We examine the effect of word alignments in Appendix D.

Method	ρ	
	TP	Loss
InputXGradient (μ)	.0588	.0756
InputXGradient (L_2)	**.7202**	**.7208**
Saliency (μ)	.5676	.5680
Saliency (L_2)	.5664	.5670
GuidedBackprop (μ)	.0026	.0020
GuidedBackprop (L_2)	.5664	.5670
IntegratedGrads (μ)	.1878	.2439
IntegratedGrads (L_2)	.6095	.5636
Activation (μ)	.5552	.5552
Activation (L_2)	.6965	.6965
LIME	.0421	.0677
Occlusion	.1480	.2049
Shapley	.2283	.2742

Table 1: Cross-lingual faithfulness results: Average correlations measured for different attribution methods on the XNLI dataset. Scores are averaged across all models including different architectures and seeds. Attributions are performed with respect to the top prediction (TP) score and the loss. InputXGradient with L_2 aggregation is the best performing method in both cases.
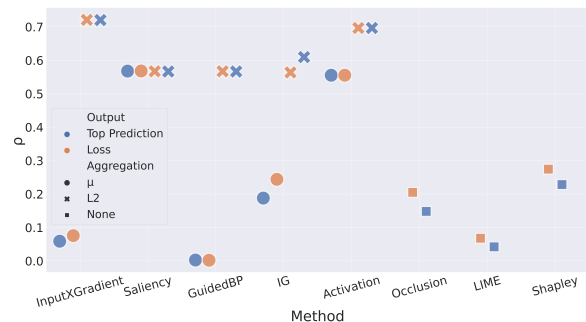

Figure 2: Comparison of cross-lingual faithfulness along output and aggregation dimensions. L_2 mostly outperforms mean (μ) aggregation and calculations with respect to the loss are the same as or slightly better than ones with respect to the top prediction score.

We fine-tune mBERT and XLM-R$_{\text{base}}$ for English on the MNLI dataset (Williams et al., 2018a) with 3 different seeds for each. For cross-lingual faithfulness evaluation, we only use the languages that are common in the top-5 languages for both types of cross-lingual models(when performing zero-shot prediction on non-English languages). This gives Bulgarian, German, Spanish and French (C = 5). The cross-lingual performance of our models on all XNLI languages appears in Appendix A.

3.2.1 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. InputXGradient with L_2 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).

Figure 2 shows the effect of aggregation methods and output mechanisms on cross-lingual faithfulness. In all cases, L_2 aggregation outperforms mean aggregation by large margins, except for Saliency, where mean aggregation is slightly better than L_2 aggregation. Since Saliency returns the absolute value, which is analogous to L_1 aggregation, the exception in the results makes sense. Considering output mechanisms, attribution scores calculated with respect to loss are more faithful than ones calculated with respect to the top prediction score in almost all cases. For Integrated Gradients with L_2 aggregation and GuidedBackprop with mean aggregation, calculating attribution scores with respect to the top prediction score performs better.

Recall that our cross-lingual faithfulness measure averages correlations across languages (Eq. 4). To analyze the effect of languages, especially the ones that are poorly represented by multilingual models, we repeat the same experiments with the worst-performing 3 languages: Thai, Swahili, and Urdu. Table 2 shows correlations per language when averaged across all combinations of methods, outputs and aggregations. The results show little variation across top-performing languages. When the relation between NLI performance and faithfulness is considered, it turns out there is a strong correlation between them (Pearson correlation coefficient and p-value are as follows: $r = 0.83$, $p = 0.02$) and poorly represented languages yield lower faithfulness scores. Detailed results per language and attribution method are given in Appendix G.

	bg	de	es	fr	th	sw	ur
ρ	.36	.38	.41	.40	.14	.27	.25
Acc	.73	.74	.77	.76	.63	.58	.62

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

Method	comp. ↑		suff. ↓	
	TP	Loss	TP	Loss
InputXGradient (μ)	.2945	.3072	.2812	.2784
InputXGradient (L_2)	.3146	.2980	**.2479**	.2682
Saliency (μ)	.3075	.3017	.2588	.2584
Saliency (L_2)	**.3158**	.3010	.2640	.2642
GuidedBackprop (μ)	.2845	.2851	.2739	.2902
GuidedBackprop (L_2)	**.3158**	.3010	.2640	.2642
IntegratedGrads (μ)	.3043	.2931	.2860	**.2308**
IntegratedGrads (L_2)	.3098	**.3160**	.2670	.2800
Activation (μ)	.2781	.2781	.2551	.2551
Activation (L_2)	.3111	.3111	.3209	.3209
LIME	.2968	.3034	.2888	.2961
Occlusion	.2898	.3080	.2887	.2656
Shapley	.2908	.3113	.2788	.2592

Table 3: Erasure-based faithfulness results: Average AOPC comprehensiveness and sufficiency scores for different attribution methods on the English split of XNLI. The scores are averaged across all models including different architectures and seeds. Attribution calculations are performed with respect to the top prediction score (TP) and the loss. Different attribution methods perform best for different output mechanisms in terms of comprehensiveness and sufficiency.

3.2.2 Erasure-based Faithfulness Experiments

Table 3 shows the results of erasure-based faithfulness evaluation (comprehensiveness and sufficiency), for each attribution method. In terms of comprehensiveness, Saliency and GuidedBackpropagation with L_2 aggregation are the most faithful attribution methods when the output is the top prediction score; IntegratedGradients with L_2 aggregation is the most faithful one when the output is the loss. For sufficiency, InputXGradient with L_2 and IntegratedGradients with mean aggregation seem to be the most faithful method for cases when the output is the top prediction score and loss, respectively. 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 beats mean aggregation
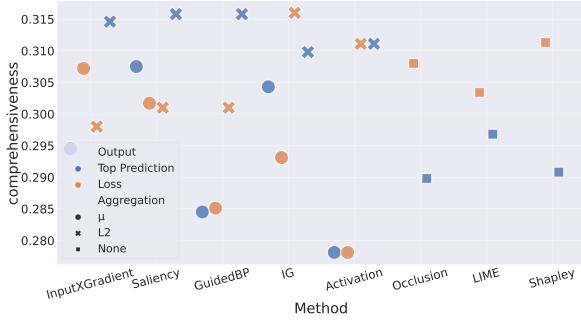
Figure 3: Comparison of comprehensiveness results along output and aggregation dimensions (higher is better). L_2 outperforms mean aggregation for most cases and calculations with respect to the loss outperform or are on par with calculations with respect to the top prediction score for non-gradient-based attribution methods.
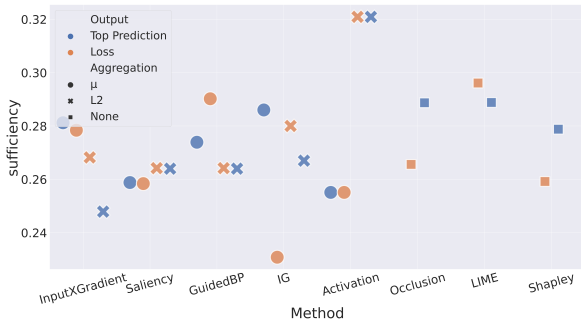

Figure 4: Comparison of sufficiency results along output and aggregation dimensions (lower is better). Different aggregation and different output mechanisms perform better for different attribution methods.

except for Saliency and InputXGradient with loss as output. While different output mechanisms are better for different methods, calculating attributions with respect to loss is as good as or slightly better than calculating with respect to the top prediction score for all non-gradient-based methods.

Figure 4 shows the effect of the aggregation method and output mechanism on sufficiency. Unlike comprehensiveness, there is no clear supremacy of one method over another for either aggregation methods or output mechanisms.

3.2.3 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 0.25 for cross-lingual faithfulness, but only 0.01 and 0.02 for comprehensiveness and sufficiency, respectively.

4 Plausibility

In this section, we present details about plausibility evaluation and results, and introduce a new dataset containing highlight-based explanations in multiple languages.

4.1 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 $\omega^{(i)}$ with gold rationales, $\mathbf{w}^{(i)}$, where $\omega^{(i)}$ 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:

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

Note that AP is the weighted mean of precisions at each threshold where the weight is the change in recall between two successive thresholds.

4.2 Plausibility Experiments

Experimental Setup We use the e-SNLI dataset (Camburu et al., 2018) to obtain human annotated highlights. As the classifier, we use a BERT-base model fine-tuned on the SNLI dataset (Bowman et al., 2015b) with 2 different seeds, as well as the one provided by TextAttack (Morris et al., 2020).

Method	MAP	
	TP	Loss
InputXGradient (μ)	.395	.397
InputXGradient (L_2)	.651	.653
Saliency (μ)	**.653**	**.655**
Saliency (L_2)	**.653**	**.655**
GuidedBackprop (μ)	.413	.414
GuidedBackprop (L_2)	**.653**	**.655**
IntegratedGrads (μ)	.473	.465
IntegratedGrads (L_2)	.633	.599
Activation (μ)	.230	.230
Activation (L_2)	.437	.437
LIME	.407	.400
Occlusion	.547	.476
Shapley	.522	.460

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

Results Table 4 shows GuidedBackprop with L_2 aggregation and Saliency with both aggregations are the most plausible methods for both types of output. Like cross-lingual faithfulness results, gradient-based methods mostly 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 scores for mean aggregation are the same as those for L_2 aggregation. Considering that Saliency returns the absolute value, which is analogous to L_1 aggregation, the exception in the results makes sense as in the cross-lingual faithfulness results. In almost all cases, calculating attribution scores with respect to loss is the same or slightly better than calculating with respect to the top prediction score. For Integrated Gradients, Occlusion, and LIME, choosing the top prediction score as output outperforms the loss.

4.3 e-XNLI dataset

Since prior studies for plausibility evaluation are limited to English-only datasets for the NLI task, we augment the multilingual XNLI dataset (Conneau et al., 2018) with highlight-based explanations
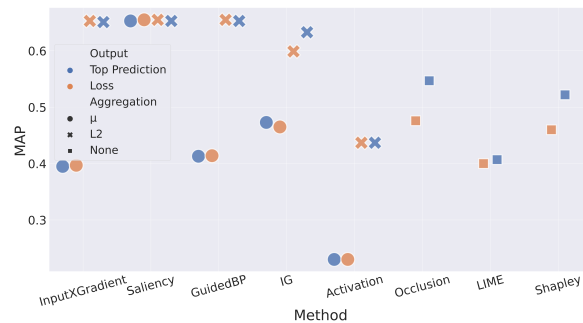

Figure 5: Comparison of plausibility results along output and aggregation dimensions. L_2 outperforms mean aggregation for almost all attribution methods and choosing loss as output is mostly the same or slightly better than the top prediction score.

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).

by utilizing attribution methods.

First, we compute attribution scores on the English split of the XNLI dataset using an mBERT model fine-tuned on MNLI and Saliency with L_2 aggregation and loss as output, which gave the most plausible attribution on e-SNLI (Section 4.2). To extract rationales from the English split, we binarize the attribution scores with respect to the threshold, 0.167, giving the best F1 score on e-SNLI with the TextAttack model.[5] Finally, we project extracted rationales to other languages using `awesome-align`.

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.

---

[5]Since there are no human-annotated highlights available for the English split of the XNLI, we first extract rationales by using the attribution method that perfromed best on a similar dataset having highlights, e-SNLI.

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 highlights 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 present more details about the human evaluation in Appendix H.

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

5 Limitations

In this work, we examine a wide range of attribution methods along output and aggregation dimensions. Prior work (Madsen et al., 2021) shows that faithfulness of attribution methods depends on both tasks and models, but our work is limited to the NLI task while considering different models. Despite the importance of NLI for evaluation in NLP (Poliak, 2020), our conclusions might not generalize to other tasks. In addition, while we experiment with multiple random seeds, our experiments are limited to two architectures: BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020).

The results of cross-lingual faithfulness experiments are sensitive to language choice as discussed in Section 3.2.1, so we present the results calculated with the languages well-represented by multilingual models. The 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 performed a human evaluation on a small subset of the dataset. Especially, training self-explanatory models with this dataset can lead to undesired outcomes, such as poor explanation quality.

6 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 for the NLI task. 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.

- To obtain the most plausible explanations, one should choose Guided Backpropagation with L_2 and Saliency with either aggregation method, and calculate scores with respect to the loss.

- To obtain the most faithful explanations, one should choose InputXGradient with L_2 regardless of output mechanism.

Finally, we present e-XNLI, a multilingual dataset with automatically generated highlight explanations, to facilitate multilingual exNLP studies.

Acknowledgements

References

Julius Adebayo, Justin Gilmer, Ian J. Goodfellow, and Been Kim. 2018. Local explanation methods for deep neural networks lack sensitivity to parameter values. *ArXiv*, abs/1810.03307.

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.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 628–635, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *EMNLP*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015b. 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*.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

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.

The Fracas Consortium, Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the framework.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *MLCW*.

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.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

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.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2021. Evaluating the faithfulness of importance measures in nlp by recursively masking allegedly important tokens and retraining. *ArXiv*, abs/2110.08412.

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.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

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 pre-training.

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.

Stefan Schweter. 2020. Berturk - bert models for turkish.

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*.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *ArXiv*, abs/1706.03825.

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*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@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. 2018a. 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.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018b. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*.

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 multilingual classifiers

Table 7 shows the average accuracies of the mBERT and XLM-R models fine-tuned on MNLI for each language in the XNLI dataset. Both models are fine-tuned for 3 epochs with learning rate 2e-5 , total batch size of 256 and 3 different seeds.

Language	mBERT	XLM-R$_{\text{base}}$
ar	0.6574	0.7132
bg	0.6952	0.7745
de	0.7120	0.7649
el	0.6724	0.7597
en	**0.8147**	**0.8436**
es	0.7504	0.7887
fr	0.7358	0.7774
hi	0.6061	0.6959
ru	0.6906	0.7549
sw	0.5137	0.6558
th	0.5468	0.7143
tr	0.6323	0.7269
ur	0.5856	0.6528
vi	0.7043	0.7466
zh	0.6952	0.7318

Table 7: Accuracies averaged across seeds of the mBERT and XLM-R$_{\text{base}}$ models fine-tuned on MNLI 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 prediction score; (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 prediction score 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 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

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

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:

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

where d is the number of dimensions for the word embedding. Similarly, using L_2 aggregation, we obtain

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

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

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

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

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

(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 $\{50, 75, 100\}$, we choose $m = 100$ for mean aggregation on calculations with respect to top prediction and $m = 50$ for all other combinations of aggregation methods and output mechanisms. 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 with XLM-R$_{\text{base}}$ and mBERT except calculations involving mean aggregation on mBERT, 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 Representational Similarity of Translation Pairs

Our cross-lingual faithfulness strategy relies on the assumption that translation pairs constitute similar inputs for a multilingual model. To test our assumption, we create a setup comparing representational similarities of inputs. First, we take premise-hypothesis pairs and their translations from XNLI dataset for the selected language pair. We encode each pair by obtaining the last hidden state representations before the classifier head. We take n representations from the source language and the corresponding representations from the target language to create source and target batch pairs, namely (b_s, b_t). Then, we create k random batches, b_i, by selecting n representations among target representations for each one and we measure the CKA similarity (Kornblith et al., 2019) of representation
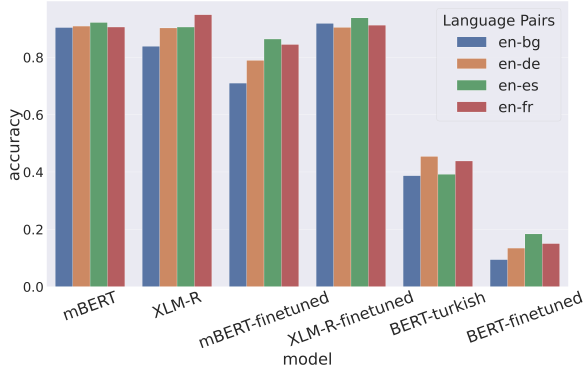
Figure 6: Accuracies for CKA similarity analysis for different models. XLM-R-finetuned and mBERT-finetuned results are averaged across models fine-tuned with different seeds for each.

batch in the source language with each batch of representation in the target language. For the sake of our assumption, we expect matching representation batches to be more similar than any batch pairs. For each batch in the source langauge, we test whether the CKA similarity measure assigns the highest similarity to the matching batches or not and compute the accuracy over batches.

We use 5000 examples from the test split of the XNLI dataset by selecting $n = 8$ and $k = 10$.

Figure 6 shows accuracies for different models. We perform our similarity analysis on multilingual models, which are vanilla mBERT, mBERT and XLM-R$_{base}$ models fine-tuned on MNLI and used in our faithfulness experiments, and monolingual models, which are BERT$_{base}$ fine-tuned on MNLI and a Turkish BERT (Schweter, 2020). The results show that translation pairs form similar inputs for multilingual models compared to monolingual models regardless of being fine-tuned. While the accuracies of fine-tuned mBERT are lower than standard mBERT, it differs among language pairs for XLM-R$_{base}$ case. Although monolingual representations of translation pairs lead to the lowest accuracies as expected, higher accuracies of Turkish BERT, which is pre-trained on a completely unrelated language, compared to fine-tuned English BERT$_{base}$ need further investigation.

D Ablation Study on Cross-lingual Faithfulness

To investigate the effect of word alignments, we run our cross-lingual faithfulness evaluation framework with random word alignments for a set of attribution methods and compare the results with the ones obtained with awesome-align (Dou and
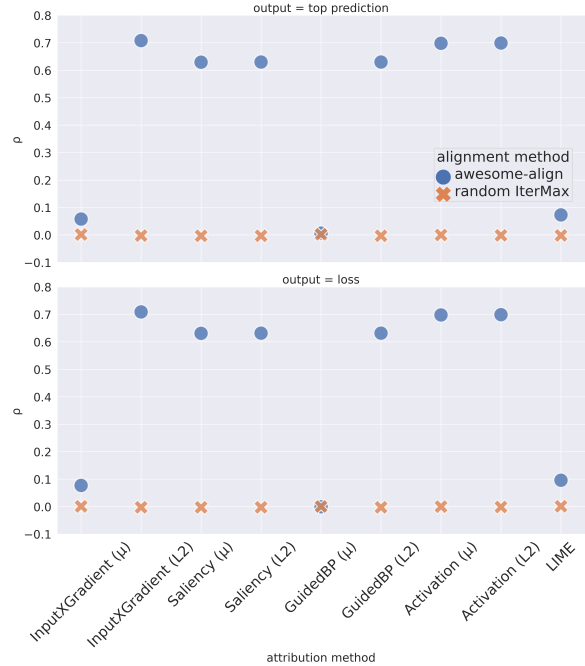
Figure 7: Comparison of cross-lingual faithfulness scores that are calculated with awesome-align and random word alignments for different attribution methods

Neubig, 2021). To obtain random alignments between translation pairs, we modify the IterMax algorithm, which Dou and Neubig (2021) proposed as a baseline method, by replacing the similarity matrix with a random matrix. We perform both types of evaluations with one of the mBERT models we fine-tuned.

Figure 7 shows the comparison of awesome-align with random word alignments. While using awesome-align provides comparable scores across attribution methods, random alignments lead to near-zero correlations (ρ). Thus we empirically show that word alignment forms a significant part of our method.

E Cross-lingual Faithfulness Results per Architecture

Table 8 shows cross-lingual faithfulness results for each architecture, mBERT and XLM-R$_{base}$, separately.

F Erasure-based Faithfulness Results per Architecture

Table 9 and Table 10 show comprehensiveness and sufficiency scores for each architecture, mBERT and XLM-R$_{base}$, separately.

Method	ρ			
	mBERT		**XLM-R$_{\text{base}}$**	
	TP	**Loss**	**TP**	**Loss**
InputXGradient (μ)	.0562 ± .002	.0758 ± .002	.0615 ± .001	.0754 ± .003
InputXGradient (L_2)	.7067 ± .001	.7078 ± .001	.7336 ± .003	.7338 ± .003
Saliency (μ)	.6269 ± .003	.6283 ± .003	.5082 ± .001	.5078 ± .002
Saliency (L_2)	.6276 ± .003	.6290 ± .003	.5053 ± .001	.5050 ± .002
GuidedBackprop (μ)	.0024 ± .003	-.0000 ± .001	.0028 ± .000	.0041 ± .001
GuidedBackprop (L_2)	.6276 ± .003	.6290 ± .003	.5053 ± .001	.5050 ± .002
IntegratedGrads (μ)	.1860 ± .008	.2680 ± .007	.1897 ± .021	.2198 ± .008
IntegratedGrads (L_2)	.5910 ± .009	.5302 ± .005	.6279 ± .018	.5970 ± .017
Activation (μ)	.6974 ± .001	.6974 ± .001	.4130 ± .001	.4130 ± .001
Activation (L_2)	.6992 ± .000	.6992 ± .000	.6938 ± .000	.6938 ± .000
LIME	.0659 ± .014	.0934 ± .006	.0182 ± .009	.0420 ± .008
Occlusion	.2281 ± .007	.3132 ± .005	.0680 ± .028	.0966 ± .007
Shapley	.3734 ± .049	.4058 ± .040	.0833 ± .016	.1426 ± .033

Table 8: Cross-lingual faithfulness results: Scores are measured for different attribution methods on the XNLI dataset and averaged across models trained with different seeds for each architecture. Attribution calculations are performed with respect to the top prediction (TP) score and the loss.

G Cross-lingual Faithfulness Results per Language

Our cross-lingual faithfulness evaluation averages correlations across languages. For completeness, we provide in Tables 11–17 the results of cross-lingual faithfulness evaluation per language.

H 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 guidelines specified in Section 3 of Camburu et al. (2018) for evaluating whether automatically-extracted highlight-based explanations are correct. Note that incorrectly predicted examples are ignored during the evaluation.

I Computational Resources

We mainly used 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 | comprehensiveness ↑ | | | |
| | mBERT | | XLM-R$_{base}$ | |
	TP	Loss	TP	Loss
InputXGradient (μ)	$.2658 \pm .016$	$.2959 \pm .012$	$.3232 \pm .011$	$.3186 \pm .031$
InputXGradient (L_2)	$.3136 \pm .011$	$.3080 \pm .005$	$.3155 \pm .021$	$.2880 \pm .005$
Saliency (μ)	$.3009 \pm .018$	$.2891 \pm .036$	$.3141 \pm .008$	$.3142 \pm .005$
Saliency (L_2)	$.3128 \pm .018$	$.2896 \pm .037$	$.3188 \pm .009$	$.3123 \pm .004$
GuidedBackprop (μ)	$.2709 \pm .002$	$.2514 \pm .039$	$.2981 \pm .023$	$.3187 \pm .015$
GuidedBackprop (L_2)	$.3128 \pm .018$	$.2896 \pm .037$	$.3188 \pm .009$	$.3123 \pm .004$
IntegratedGrads (μ)	$.2557 \pm .033$	$.2618 \pm .010$	$.3529 \pm .021$	$.3244 \pm .018$
IntegratedGrads (L_2)	$.2989 \pm .004$	$.2969 \pm .014$	$.3208 \pm .028$	$.3350 \pm .031$
Activation (μ)	$.2504 \pm .009$	$.2504 \pm .009$	$.3057 \pm .006$	$.3057 \pm .006$
Activation (L_2)	$.2940 \pm .010$	$.2940 \pm .010$	$.3282 \pm .017$	$.3282 \pm .017$
LIME	$.2733 \pm .026$	$.2657 \pm .016$	$.3203 \pm .028$	$.3412 \pm .024$
Occlusion	$.2727 \pm .034$	$.3101 \pm .014$	$.3068 \pm .029$	$.3060 \pm .016$
Shapley	$.2660 \pm .032$	$.3123 \pm .007$	$.3157 \pm .019$	$.3103 \pm .023$

Table 9: Comprehensiveness scores per architecture on the English split of XNLI dataset: Scores are measured for different attribution methods on the XNLI dataset and averaged across models trained with different seeds for each architecture. Attribution calculations are performed with respect to the top prediction (TP) score and the loss.

| Method | sufficiency ↓ | | | |
| | mBERT | | XLM-R$_{base}$ | |
	TP	Loss	TP	Loss
InputXGradient (μ)	$.2812 \pm .013$	$.2716 \pm .021$	$.2812 \pm .007$	$.2852 \pm .014$
InputXGradient (L_2)	$.2616 \pm .043$	$.2684 \pm .027$	$.2342 \pm .023$	$.2681 \pm .026$
Saliency (μ)	$.2451 \pm .028$	$.2613 \pm .029$	$.2724 \pm .011$	$.2555 \pm .004$
Saliency (L_2)	$.2477 \pm .022$	$.2629 \pm .027$	$.2804 \pm .008$	$.2654 \pm .007$
GuidedBackprop (μ)	$.2637 \pm .031$	$.2913 \pm .007$	$.2841 \pm .032$	$.2891 \pm .023$
GuidedBackprop (L_2)	$.2477 \pm .022$	$.2629 \pm .027$	$.2804 \pm .008$	$.2654 \pm .007$
IntegratedGrads (μ)	$.2985 \pm .008$	$.2471 \pm .024$	$.2734 \pm .020$	$.2145 \pm .011$
IntegratedGrads (L_2)	$.2784 \pm .010$	$.2788 \pm .021$	$.2556 \pm .006$	$.2812 \pm .025$
Activation (μ)	$.2024 \pm .017$	$.2024 \pm .017$	$.3079 \pm .010$	$.3079 \pm .010$
Activation (L_2)	$.3340 \pm .003$	$.3340 \pm .003$	$.3078 \pm .005$	$.3078 \pm .005$
LIME	$.2610 \pm .005$	$.2610 \pm .012$	$.3167 \pm .048$	$.3311 \pm .005$
Occlusion	$.2820 \pm .006$	$.2475 \pm .008$	$.2955 \pm .015$	$.2837 \pm .006$
Shapley	$.2538 \pm .008$	$.1967 \pm .008$	$.3037 \pm .035$	$.3218 \pm .013$

Table 10: Sufficiency scores per architecture on the English split of XNLI dataset: Scores are measured for different attribution methods on the XNLI dataset and averaged across models trained with different seeds for each architecture. Attribution calculations are performed with respect to the top prediction (TP) score and the loss.

| | ρ | | | |
| Method | mBERT | | XLM-R$_{base}$ | |
	TP	Loss	TP	Loss
InputXGradient (μ)	.0302 ± .001	.0406 ± .001	.0506 ± .003	.0661 ± .002
InputXGradient (L_2)	.6731 ± .001	.6741 ± .002	.7188 ± .003	.7189 ± .003
Saliency (μ)	.5778 ± .004	.5793 ± .004	.4954 ± .000	.4951 ± .002
Saliency (L_2)	.5787 ± .004	.5803 ± .004	.4935 ± .000	.4930 ± .001
GuidedBackprop (μ)	.0015 ± .001	-.0045 ± .003	.0020 ± .001	.0023 ± .004
GuidedBackprop (L_2)	.5787 ± .004	.5803 ± .004	.4935 ± .000	.4930 ± .001
IntegratedGrads (μ)	.1248 ± .005	.2003 ± .002	.1768 ± .021	.2037 ± .009
IntegratedGrads (L_2)	.5287 ± .014	.4585 ± .011	.6165 ± .016	.5827 ± .016
Activation (μ)	.6080 ± .001	.6080 ± .001	.3824 ± .001	.3824 ± .001
Activation (L_2)	.6653 ± .000	.6653 ± .000	.6825 ± .000	.6825 ± .000
LIME	.0561 ± .015	.0803 ± .008	.0115 ± .007	.0359 ± .009
Occlusion	.1635 ± .008	.2395 ± .004	.0509 ± .021	.0754 ± .004
Shapley	.3348 ± .055	.3639 ± .044	.0649 ± .016	.1165 ± .032

Table 11: Cross-lingual faithfulness results for the Bulgarian split of XNLI dataset: Scores are measured for different attribution methods on the XNLI dataset and averaged across models trained with different seeds for each architecture. Attribution calculations are performed with respect to the top prediction (TP) score and the loss.

| | ρ | | | |
| Method | mBERT | | XLM-R$_{base}$ | |
	TP	Loss	TP	Loss
InputXGradient (μ)	.0493 ± .003	.0717 ± .004	.0621 ± .001	.0752 ± .003
InputXGradient (L_2)	.7052 ± .001	.7067 ± .001	.7321 ± .003	.7329 ± .003
Saliency (μ)	.6152 ± .003	.6168 ± .003	.4936 ± .002	.4929 ± .004
Saliency (L_2)	.6159 ± .003	.6175 ± .003	.4906 ± .002	.4900 ± .003
GuidedBackprop (μ)	.0041 ± .005	.0008 ± .002	.0030 ± .001	.0010 ± .001
GuidedBackprop (L_2)	.6159 ± .003	.6175 ± .003	.4906 ± .002	.4900 ± .003
IntegratedGrads (μ)	.1919 ± .011	.2788 ± .011	.1814 ± .019	.2219 ± .008
IntegratedGrads (L_2)	.5935 ± .007	.5361 ± .005	.6245 ± .018	.5930 ± .017
Activation (μ)	.6960 ± .000	.6960 ± .000	.4115 ± .001	.4115 ± .001
Activation (L_2)	.7012 ± .000	.7012 ± .000	.6934 ± .000	.6934 ± .000
LIME	.0692 ± .012	.0955 ± .005	.0186 ± .009	.0399 ± .008
Occlusion	.2226 ± .006	.3117 ± .005	.0695 ± .028	.0978 ± .005
Shapley	.3843 ± .046	.4145 ± .036	.0840 ± .017	.1381 ± .030

Table 12: Cross-lingual faithfulness results for the German split of XNLI dataset: Scores are measured for different attribution methods on the XNLI dataset and averaged across models trained with different seeds for each architecture. Attribution calculations are performed with respect to the top prediction (TP) score and the loss.

| Method | ρ | | | |
| | mBERT | | XLM-R$_{\text{base}}$ | |
	TP	Loss	TP	Loss
InputXGradient (μ)	$.0747 \pm .001$	$.1009 \pm .003$	$.0726 \pm .003$	$.0856 \pm .005$
InputXGradient (L_2)	$.7179 \pm .001$	$.7186 \pm .001$	$.7460 \pm .003$	$.7461 \pm .003$
Saliency (μ)	$.6576 \pm .002$	$.6592 \pm .002$	$.5205 \pm .001$	$.5205 \pm .002$
Saliency (L_2)	$.6582 \pm .002$	$.6598 \pm .002$	$.5170 \pm .001$	$.5172 \pm .001$
GuidedBackprop (μ)	$.0004 \pm .003$	$.0026 \pm .003$	$.0035 \pm .003$	$.0049 \pm .002$
GuidedBackprop (L_2)	$.6582 \pm .002$	$.6598 \pm .002$	$.5170 \pm .001$	$.5172 \pm .001$
IntegratedGrads (μ)	$.2189 \pm .008$	$.3041 \pm .005$	$.2037 \pm .020$	$.2374 \pm .006$
IntegratedGrads (L_2)	$.6145 \pm .010$	$.5584 \pm .006$	$.6384 \pm .018$	$.6083 \pm .018$
Activation (μ)	$.7521 \pm .001$	$.7521 \pm .001$	$.4311 \pm .001$	$.4311 \pm .001$
Activation (L_2)	$.7071 \pm .000$	$.7071 \pm .000$	$.7135 \pm .000$	$.7135 \pm .000$
LIME	$.0716 \pm .013$	$.1032 \pm .009$	$.0259 \pm .012$	$.0500 \pm .006$
Occlusion	$.2693 \pm .008$	$.3594 \pm .006$	$.0839 \pm .035$	$.1202 \pm .014$
Shapley	$.3928 \pm .047$	$.4238 \pm .039$	$.1041 \pm .022$	$.1716 \pm .035$

Table 13: Cross-lingual faithfulness results for the Spanish split of XNLI dataset: Scores are measured for different attribution methods on the XNLI dataset and averaged across models trained with different seeds for each architecture. Attribution calculations are performed with respect to the top prediction (TP) score and the loss.

| Method | ρ | | | |
| | mBERT | | XLM-R$_{\text{base}}$ | |
	TP	Loss	TP	Loss
InputXGradient (μ)	$.0707 \pm .003$	$.0902 \pm .003$	$.0605 \pm .001$	$.0746 \pm .001$
InputXGradient (L_2)	$.7308 \pm .001$	$.7317 \pm .001$	$.7374 \pm .003$	$.7372 \pm .003$
Saliency (μ)	$.6570 \pm .003$	$.6578 \pm .002$	$.5234 \pm .001$	$.5226 \pm .002$
Saliency (L_2)	$.6574 \pm .003$	$.6583 \pm .002$	$.5202 \pm .001$	$.5199 \pm .002$
GuidedBackprop (μ)	$.0034 \pm .005$	$.0010 \pm .003$	$.0028 \pm .000$	$.0082 \pm .002$
GuidedBackprop (L_2)	$.6574 \pm .003$	$.6583 \pm .002$	$.5202 \pm .001$	$.5199 \pm .002$
IntegratedGrads (μ)	$.2082 \pm .009$	$.2887 \pm .009$	$.1968 \pm .025$	$.2163 \pm .009$
IntegratedGrads (L_2)	$.6274 \pm .008$	$.5676 \pm .004$	$.6321 \pm .017$	$.6040 \pm .016$
Activation (μ)	$.7333 \pm .001$	$.7333 \pm .001$	$.4271 \pm .001$	$.4271 \pm .001$
Activation (L_2)	$.7234 \pm .000$	$.7234 \pm .000$	$.6857 \pm .000$	$.6857 \pm .000$
LIME	$.0668 \pm .016$	$.0945 \pm .005$	$.0168 \pm .011$	$.0420 \pm .008$
Occlusion	$.2568 \pm .008$	$.3422 \pm .007$	$.0678 \pm .029$	$.0929 \pm .007$
Shapley	$.3816 \pm .047$	$.4209 \pm .040$	$.0803 \pm .011$	$.1442 \pm .034$

Table 14: Cross-lingual faithfulness results for the French split of XNLI dataset: Scores are measured for different attribution methods on the XNLI dataset and averaged across models trained with different seeds for each architecture. Attribution calculations are performed with respect to the top prediction (TP) score and the loss.

| Method | ρ | | | |
| | mBERT | | XLM-R | |
	TP	Loss	TP	Loss
InputXGradient (μ)	.0024 ± .001	.0051 ± .001	.0076 ± .002	.0174 ± .001
InputXGradient (L_2)	.2924 ± .001	.2953 ± .001	.2749 ± .001	.2761 ± .001
Saliency (μ)	.2166 ± .003	.2196 ± .004	.1712 ± .002	.1702 ± .002
Saliency (L_2)	.2168 ± .003	.2198 ± .004	.1705 ± .002	.1692 ± .002
GuidedBackprop (μ)	.0007 ± .003	.0007 ± .001	.0017 ± .001	.0025 ± .000
GuidedBackprop (L_2)	.2168 ± .003	.2198 ± .004	.1705 ± .002	.1692 ± .002
IntegratedGrads (μ)	.0286 ± .008	.0758 ± .012	.0963 ± .014	.1173 ± .017
IntegratedGrads (L_2)	.2597 ± .009	.2433 ± .009	.2059 ± .008	.1970 ± .008
Activation (μ)	.2462 ± .000	.2462 ± .000	.1307 ± .000	.1307 ± .000
Activation (L_2)	.2127 ± .000	.2127 ± .000	.2007 ± .000	.2007 ± .000
LIME	.0281 ± .022	.0173 ± .008	.0041 ± .002	.1552 ± .003
Occlusion	.0451 ± .011	.0591 ± .009	.0128 ± .003	.0305 ± .002
Shapley	.3001 ± .063	.2461 ± .051	.0283 ± .013	.0741 ± .015

Table 15: Cross-lingual faithfulness results for the Thai split of XNLI dataset: Scores are measured for different attribution methods on the XNLI dataset and averaged across models trained with different seeds for each architecture. Attribution calculations are performed with respect to the top prediction (TP) class and the loss.

| Method | ρ | | | |
| | mBERT | | XLM-R | |
	TP	Loss	TP	Loss
InputXGradient (μ)	.0064 ± .001	.0132 ± .002	.0167 ± .002	.0253 ± .003
InputXGradient (L_2)	.4598 ± .001	.4616 ± .001	.5254 ± .002	.5255 ± .002
Saliency (μ)	.4083 ± .002	.4107 ± .002	.4136 ± .002	.4132 ± .002
Saliency (L_2)	.4084 ± .002	.4108 ± .002	.4110 ± .002	.4104 ± .002
GuidedBackprop (μ)	.0043 ± .002	-.0015 ± .003	.0002 ± .003	.0015 ± .001
GuidedBackprop (L_2)	.4084 ± .002	.4108 ± .002	.4110 ± .002	.4104 ± .002
IntegratedGrads (μ)	.0713 ± .005	.1239 ± .006	.1137 ± .021	.1223 ± .005
IntegratedGrads (L_2)	.3806 ± .005	.3298 ± .005	.4883 ± .009	.4634 ± .011
Activation (μ)	.4686 ± .001	.4686 ± .001	.3400 ± .001	.3400 ± .001
Activation (L_2)	.4987 ± .000	.4987 ± .000	.5449 ± .000	.5449 ± .000
LIME	.0257 ± .015	.0752 ± .002	.0128 ± .005	.1854 ± .001
Occlusion	.0537 ± .003	.0988 ± .001	.0416 ± .016	.0628 ± .010
Shapley	.2424 ± .044	.2622 ± .044	.0612 ± .015	.1174 ± .014

Table 16: Cross-lingual faithfulness results for the Swahili split of XNLI dataset: Scores are measured for different attribution methods on the XNLI dataset and averaged across models trained with different seeds for each architecture. Attribution calculations are performed with respect to the top prediction (TP) class and the loss.

Method	ρ			
	mBERT		XLM-R	
	TP	Loss	TP	Loss
InputXGradient (μ)	$.0156 \pm .003$	$.0211 \pm .005$	$.0147 \pm .002$	$.0225 \pm .004$
InputXGradient (L_2)	$.5522 \pm .002$	$.5533 \pm .002$	$.5031 \pm .003$	$.5037 \pm .003$
Saliency (μ)	$.4492 \pm .004$	$.4509 \pm .004$	$.2512 \pm .004$	$.2513 \pm .004$
Saliency (L_2)	$.4499 \pm .004$	$.4515 \pm .004$	$.2495 \pm .004$	$.2496 \pm .004$
GuidedBackprop (μ)	$-.0003 \pm .003$	$.0006 \pm .001$	$-.0017 \pm .003$	$.0011 \pm .006$
GuidedBackprop (L_2)	$.4499 \pm .004$	$.4515 \pm .004$	$.2495 \pm .004$	$.2496 \pm .004$
IntegratedGrads (μ)	$.0700 \pm .003$	$.1453 \pm .003$	$.0886 \pm .009$	$.1187 \pm .008$
IntegratedGrads (L_2)	$.4451 \pm .012$	$.3909 \pm .008$	$.4168 \pm .012$	$.3955 \pm .014$
Activation (μ)	$.4700 \pm .000$	$.4700 \pm .000$	$.2407 \pm .001$	$.2407 \pm .001$
Activation (L_2)	$.5688 \pm .000$	$.5688 \pm .000$	$.4820 \pm .000$	$.4820 \pm .000$
LIME	$.0398 \pm .013$	$.0593 \pm .003$	$.0049 \pm .003$	$.1077 \pm .006$
Occlusion	$.0815 \pm .005$	$.1382 \pm .006$	$.0037 \pm .012$	$.0213 \pm .013$
Shapley	$.2557 \pm .044$	$.2915 \pm .038$	$.0229 \pm .008$	$.0490 \pm .009$

Table 17: Cross-lingual faithfulness results for the Urdu split of XNLI dataset: Scores are measured for different attribution methods on the XNLI dataset and averaged across models trained with different seeds for each architecture. Attribution calculations are performed with respect to the top prediction (TP) class and the loss.