



UvA-DARE (Digital Academic Repository)

Feature Interactions Reveal Linguistic Structure in Language Models

Jumelet, J.; Zuidema, W.

DOI

[10.18653/v1/2023.findings-acl.554](https://doi.org/10.18653/v1/2023.findings-acl.554)

Publication date

2023

Document Version

Final published version

Published in

Findings of the Association for Computational Linguistics: ACL 2023

License

CC BY

[Link to publication](#)

Citation for published version (APA):

Jumelet, J., & Zuidema, W. (2023). Feature Interactions Reveal Linguistic Structure in Language Models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023: July 9-14, 2023* (pp. 8697–8712). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.554>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Feature Interactions Reveal Linguistic Structure in Language Models

Jaap Jumelet Willem Zuidema
Institute for Logic, Language and Computation
University of Amsterdam
{j.w.d.jumelet, w.zuidema}@uva.nl

Abstract

We study *feature interactions* in the context of *feature attribution* methods for post-hoc interpretability. In interpretability research, getting to grips with feature interactions is increasingly recognised as an important challenge, because interacting features are key to the success of neural networks. Feature interactions allow a model to build up hierarchical representations for its input, and might provide an ideal starting point for the investigation into linguistic structure in language models. However, uncovering the exact role that these interactions play is also difficult, and a diverse range of interaction attribution methods has been proposed. In this paper, we focus on the question which of these methods most *faithfully* reflects the inner workings of the target models. We work out a *grey box* methodology, in which we train models to perfection on a formal language classification task, using PCFGs. We show that under specific configurations, some methods are indeed able to uncover the grammatical rules acquired by a model. Based on these findings we extend our evaluation to a case study on language models, providing novel insights into the linguistic structure that these models have acquired.¹

1 Introduction

Feature attribution methods (FAMs) are a popular family of tools for explaining the behaviour of deep learning models, by explaining a prediction in terms of contributions of individual features (Ribeiro et al., 2016; Lundberg and Lee, 2017). There are many such methods proposed, and mathematical results (such as axiomatic approaches based on game theory) and theoretical frameworks (such as Covert et al. (2021)’s ‘Explaining by Removing’) are starting to offer a good understanding of how different methods relate to one another.

However, there are also some important shortcomings. Perhaps most importantly, popular FAMs

¹All code and data is available here: <https://github.com/jumelet/fidam-eval>

mostly ignore the existence of interactions between the effects of features on the prediction. This is problematic, because **Feature Interactions** are widely seen as a major factor in the success of neural networks (Goodfellow et al., 2016). This is all the more important in domains such as language and music processing, because feature interactions allow neural networks to model hierarchical representations of their input, which is considered a key design feature of language and music. To address these shortcomings, there is now an emerging literature on **feature interaction detection and attribution methods** (FIDAMs) that explain model predictions in terms of interacting features (Tsang et al., 2020; Janizek et al., 2021).

However, assessing the faithfulness of FIDAMs is even more challenging than assessing the faithfulness of feature attribution methods more generally (Jacovi and Goldberg, 2021). In this paper, we present a systematic framework to characterise FIDAMs, and derive several new FIDAMs based on that framework. We then proceed with creating an evaluation pipeline that measures a FIDAM’s ability to recover the structural rules for which we have good evidence that they play an important role in the target model’s performance (Figure 1). We first test this on a set of small-scale formal language tasks, that provide stronger faithfulness guarantees. Finally, we present a case study of a large language model on the CoLA task for linguistic acceptability.

We find that the performance of FIDAMs is very variable, and that the performance on the small-scale formal language tasks may not be predictive of the performance of methods on the large-scale natural language task. This is an illustration of what we call the **Attribution Generalisation problem**. We argue that this problem remains a key open problem in the study of explanation methods in general.

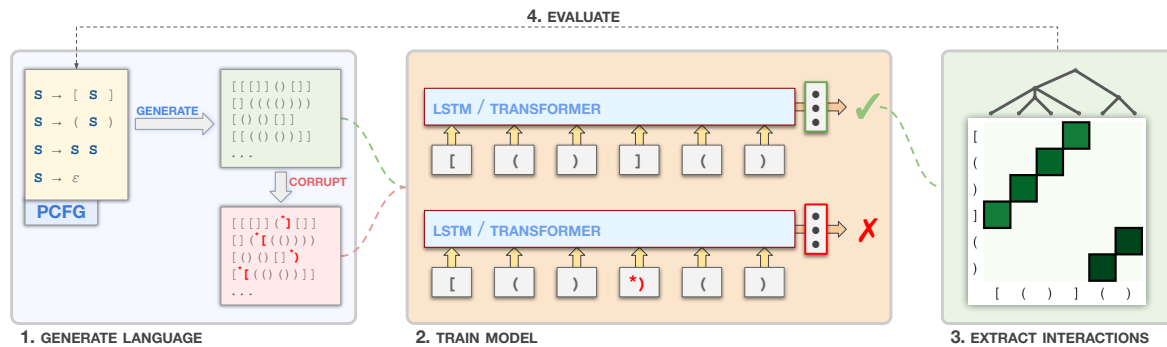


Figure 1: We generate a corpus based on a PCFG, and create negative examples by corrupting the generated corpus. Next, we train a neural model to predict whether a string is well-formed, forcing the model to obtain a comprehensive understanding of the rules of the language. Then, we extract the internal interactions by using FIDAMs described in §4, allowing us to directly evaluate the grammatical knowledge of the neural model.

2 Related Work: Assessing Faithfulness

In this section we discuss related work on assessing the faithfulness of feature attribution methods (FAMs). A model explanation ideally provides better insights into model behaviour. However, it is important that an explanation is faithful to the reasoning of the model, and not merely plausible to a researcher. Unfortunately, attribution models can yield vastly different outcomes (Neely et al., 2022).

Defining a notion of faithfulness itself is an ongoing debate, and it has been argued that we should not be aiming for a binary notion, but a graded one instead (Jacovi and Goldberg, 2021). To this end, various methodologies have been proposed to evaluate the faithfulness of explanation methods.

One research direction introduces metrics to evaluate faithfulness by quantifying the impact of features that were deemed to contribute the most by an attribution method. Hooker et al. (2019) does this by *retraining* a model on data from which the most contributing features have been removed. DeYoung et al. (2020) provide a more direct measure, by quantifying changes in model predictions when only a subset of the most contributing features is fed to model. Atanasova et al. (2020) build on this notion, introducing a range of diagnostic metrics that capture various aspects of explanation quality including faithfulness, human rationale agreement, and explanation consistency. Jain et al. (2020) ensure and evaluate faithfulness by only allowing a model access to the set of features that were deemed important by the explanation method, which has also been shown to improve model robustness (Wiegrefe et al., 2021; Ross et al., 2022).

Another line of work modifies the training data in such a way that we obtain guarantees of cer-

tain features the model must be paying attention to when making a prediction: e.g. by shuffling test data such that only part of the input resembles the statistics from the train set (Pörner et al., 2018), or by explicitly adding exploitable heuristics in the train set (Bastings et al., 2022; Adebayo et al., 2022). These two approaches could be characterised as *grey box* models: we adapt the data in such a way that we gain a degree of confidence what cues the model must be relying on, without having a full understanding of the model’s internal reasoning. A *glass box* model, on the other hand, is a model whose behaviour is fully understood: it’s not derived by training a model on a task, but hand-crafted. Hao (2020) utilises such models to evaluate FAMs on formal language tasks, providing more robust guarantees on model behaviour.

Our own approach is related to the first line of research, making use of *grey box* models. Instead of evaluating FAMs, we evaluate FIDAMs, that provide more comprehensive insights into model reasoning. Deployment of such methods within NLP has been fairly limited, and as such evaluating their faithfulness in a language context has been an underexplored research topic.

3 A Framework for Characterising FIDAMs

Feature attribution methods typically decompose a model prediction into a sum of feature contributions (Sundararajan et al., 2017; Lundberg and Lee, 2017). A large contribution then indicates that this feature played an important role in a model’s prediction. Although feature attributions can provide meaningful insights into the inner model dynamics, they paint a fairly limited picture of the model be-

haviour. Most importantly, **interactions** between features are lumped together, making it impossible to discern whether a large contribution of a feature stemmed from that feature alone, or from its interaction with neighbouring features. To address this, multiple methods have been proposed that decompose a model prediction into a sum of feature interactions, based on similar mathematical formalism as those of feature attributions.

Notation A neural network is represented as a single function f . The input to f is denoted as \mathbf{x} , which consists of N input features. A partial input \mathbf{x}_S only consists of input features $S \subseteq N$. A value function $v(\mathbf{x}_S)$ quantifies the model output on the partial input \mathbf{x}_S . Padding the missing features in \mathbf{x}_S with replacement features $\mathbf{x}'_{\setminus S}$ is denoted as $\mathbf{x}_S \cup \mathbf{x}'_{\setminus S}$. The attribution value of feature i is denoted as ϕ_i , and the interaction effect of a set of features \mathcal{I} is denoted as $\Gamma_{\mathcal{I}}$.

Attribution Dimensions Attribution methods can generally be characterised along two dimensions (Covert et al., 2021): 1) how the method deals with feature removal, and 2) how the impact of removing a feature is quantified. FIDAMs are built on the same principles as FAMs, and can be categorised along the same two dimension. By discerning these two dimensions we can separately evaluate their impact on the faithfulness of the attribution method. Furthermore, we can combine feature removal procedures with influence quantification methods in order to obtain novel attribution methods, an observation that has also been made in the context of FIDAMs by Jiang and Steinert-Threlkeld (2023), who, concurrent to our work, provide a general framework for characterising FIDAMs.

3.1 Feature Removal

It is not straight-forward to define the absence of a feature to a model’s input. The main goal here is to replace the removed feature with a neutral **baseline**, that adequately represents the absence of the feature. Methods often make use of a neutral input feature, the **static baseline** \mathbf{x}' , such as a zero-valued embedding or a pad token:

$$v(\mathbf{x}_S) = f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S}) \quad (1)$$

This may, however, lead to input that lies outside of the original input distribution (Kim et al., 2020). The reason why this is problematic is that the model

may behave erratically on such modified input, posing issues to the faithfulness of the explanation.

Instead of using a static baseline, we can also opt to use a baseline that is sampled from a *background distribution* (Datta et al., 2016). There exist two approaches to this procedure (Sundararajan and Najmi, 2020; Chen et al., 2020b). The **observational conditional expectation** samples the baseline features from a distribution that is conditioned on the set of features that are still present in the input (Frye et al., 2020; Aas et al., 2021):

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}} \left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S}) \mid \mathbf{x}_S \right] \quad (2)$$

The **interventional conditional expectation** drops the conditional, and samples the baseline features from an independent distribution:

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}} \left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S}) \right] \quad (3)$$

There exist two motivations for the latter approach: Lundberg and Lee (2017) drop the conditional expectation for computational reasons, allowing them to approximate the observational conditional expectation. Janzing et al. (2020) provide a perspective derived from causality theory, stating that the *intervention* of removing a feature should break the dependence between the baseline and remaining features, and hence conditioning on these features is fundamentally wrong.

The previous two methods sample baseline values for individual missing features, but we can also compute the expectation over the range of possible baselines. This yields the technique of **expected explanations** (Erion et al., 2021), in which attributions with different static baselines are averaged out over a background distribution D :

$$\phi_i = \mathbb{E}_{\mathbf{x}' \sim D} \left[\phi_i(\mathbf{x}; \mathbf{x}') \right] \quad (4)$$

3.2 Quantifying Feature Influence

The simplest method of quantifying the influence of a feature is expressed as the output difference after **ablating** the feature:

$$\phi_i = v(\mathbf{x}) - v(\mathbf{x}_{\setminus i}) \quad (5)$$

Note that this formulation can be combined with any of the feature removal methods: e.g. Occlusion (Zeiler and Fergus, 2014) combines this influence method with a static baseline (Eq. 1), whereas Kim et al. (2020) combines it with the observational

conditional expectation (Eq. 2), employing BERT as the conditional distribution.

A more involved method leverages a technique from the field of game theory, called the **Shapley value** (Shapley, 1953). Shapley values were originally introduced in the domain of cooperative games, in which players can form coalitions to change the outcome of the game. This setup can be transferred directly to machine learning models, in which features now take up the role of the players. A Shapley value expresses the contribution of a feature as the marginal gain of including that feature in the input, averaged over all possible coalitions of features.

4 FIDAMs

We now address a series of interaction methods that we use in our own experiments.

Group Ablation The feature influence principle of Equation 5 can straightforwardly be extended to *groups* of features. In our experiments we will focus on pairwise interactions, but any kind of feature subset can be used here.

$$\Gamma_{i,j} = v(\mathbf{x}) - v(\mathbf{x}_{\setminus ij}) \quad (6)$$

Archipelago Explaining model behaviour in terms of pairwise interactions will already yield a better portrayal of its internal behaviour than ‘flat’ attributions, but it neglects the interactions that occur within larger groups of features. Archipelago (Tsang et al., 2020) splits up the feature interaction procedure into two phases: first an interaction detection method is performed that clusters features into interaction sets, and afterwards interaction scores are assigned to these sets as a whole. Interaction detection is based on measuring the non-additive effect of pairs of features. The interaction effect that is assigned to an interaction set \mathcal{I} is expressed as follows, with respect to a static baseline \mathbf{x}' :

$$\Gamma_{\mathcal{I}} = f(\mathbf{x}_{\mathcal{I}} \cup \mathbf{x}'_{\setminus \mathcal{I}}) - f(\mathbf{x}') \quad (7)$$

Note that Archipelago expresses the interaction effect inversely compared to the Group Ablation procedure: instead of measuring the impact of removing a group of features, we now measure the impact of solely keeping this group in the input.

Shapley(-Taylor) Interaction Index Both the previous methods base interaction effects on direct output differences. We can modify the formulation

of the Shapley value to yield interaction effects. This modification was originally introduced in the field of game theory, called the Shapley Interaction Index (SII, Owen, 1972; Grabisch and Roubens, 1999). Instead of computing the marginal gain that is achieved by a single feature, we now compute the marginal gain of *groups* of features. The Shapley-Taylor Interaction Index (STII, Sundararajan et al., 2020) is an extension of SII, satisfying additional theoretical properties.

Hessian Analogous to utilising the gradient for feature attributions, we can employ the second-order derivative to quantify interactions between features, which is captured by the Hessian matrix. Friedman and Popescu (2008) and Sorokina et al. (2008) consider an interaction between two variables to exist when the effect of one variable on the response depends on values of the other variable, which can be expressed in terms of the second-order partial derivative:

$$\Gamma_{i,j} = \left[\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right]^2$$

A common approach when using the gradient of a model as a proxy for feature importance is to multiply it with the input embeddings (Shrikumar et al., 2017; Ancona et al., 2019): in our experiments we consider an analogous method to the Hessian that we call **Hessian \times Input**.

Integrated Hessians Directly using the Hessian as explanation method is prone to the same caveats as using the gradient: the interactions signal may vanish due to saturation. Integrated Hessians (IH, Janizek et al., 2021) address this issue by integrating over the Hessian manifold along a path between the input and a baseline. This is achieved by applying the method of Integrated Gradients (Sundararajan et al., 2017) to itself. An IH interaction between features i and j can hence be interpreted as the contribution of i to the contribution of j to the models prediction. The path integral between input and baseline is approximated via a Riemann sum interpolation.

Other Methods The methods explained thus far have all been incorporated in our experimental pipeline. The scope of our work focuses mainly on *pairwise* interactions, but methods that extract higher-order interactions have been proposed as well (Jin et al., 2020). Comparing such methods to linguistic structure is an exciting avenue that we

leave open to future work. Other interaction methods that were not considered include two methods that preceded Archipelago: Neural Interaction Detection (Tsang et al., 2018a) and MAHE (Tsang et al., 2018b). The feature attribution method Contextual Decomposition (Murdoch et al., 2018) has been extended to extract interactions as well (Singh et al., 2019; Saphra and Lopez, 2020; Chen et al., 2020a), but these methods place the constraint that only contiguous groups of features can interact. Integrated Directional Gradients (Sikdar et al., 2021), an extension of Integrated Gradients to capture *group attributions*, could be adapted to our framework, but we leave this open for future work.

5 Evaluating FIDAMs

The final component of our framework is a methodology for evaluating the faithfulness of FIDAMs. To lay a robust foundation for such work, we propose to evaluate a range of interaction methods and baselines on smaller deep learning models (using LSTM and Transformer architectures) that have been trained to recognise formal languages, based on a probabilistic context-free grammar (PCFG).

Our models are trained on a binary language classification task, in which a model needs to learn to discern between well-formed strings and minimally corrupted counterparts. Models are trained to perfection (100% accuracy) on both train and test set. To obtain perfect performance, a model must rely solely on the grammatical rules that underlie the language, without resorting to spurious heuristics, because only these results allow completely solving the task. This way, due to the controlled nature of the task, we obtain a high degree of confidence about the model’s behaviour.

The goal of our experimental approach is to recover the structure of the language *based on the trained model itself*. This is achieved by the FIDAMs outlined in §4. We aim to uncover whether a structural dependency between two features results in a high interaction effect. Since our models have been trained to perfection, this allows us to employ our setup as a way of measuring the **faithfulness** of a FIDAM. A method that assigns a high interaction effect to features that contain a dependency in the original grammar is able to provide a faithful reflection of a model’s understanding of the task. By testing a wide range of FIDAMs and baselines we can uncover which configuration yields the most faithful explanations. A graphical overview of our

approach is depicted in Figure 1.

Task The binary language classification task is set up by generating positive examples D^+ , based on some PCFG, and negative examples D^- , derived from minimally corrupting the positive examples. We split the union of these two sets into a random train/test split of 80/20%. We train our models with a default cross-entropy loss, using the AdamW optimiser (Loshchilov and Hutter, 2019), a learning rate of 0.01, and a batch size of 48.

Models Our pipeline permits the use of any kind of neural model architecture, in our experiments we considered both LSTMs (Hochreiter and Schmidhuber, 1997) and Transformers (Vaswani et al., 2017). In our experiments we report the results of the LSTM model, but we observed similar results for Transformers: due to the black-box approach of our explanation procedure the architecture itself is not of great importance. The models are deliberately small: we use an embedding size that is equal to the number of symbols in the language it is trained on, a hidden state size of 20, and a single layer. This results in models that provide a compute-friendly test bed for evaluating the FIDAMs.

Evaluation We focus on *pairwise* interactions: interactions between individual pairs of features. A FIDAM that extracts pairwise interactions for an input sequence $\mathbf{x} \in \mathbb{R}^N$ returns a matrix of interaction effects $\Gamma \in \mathbb{R}^{N \times N}$. Since our goal is to uncover whether structural dependencies result in high interaction effects, we approach the evaluation of the interaction matrix as a retrieval task. By aggregating and normalising the *rank* of each interaction of interest we can quantify the performance of a FIDAM. We call this metric the **Average Relative Rank** (ARR):

$$ARR(\Gamma, \mathcal{I}) = \frac{1}{|\mathcal{I}|} \sum_{i,j \in \mathcal{I}} \frac{R(\Gamma_i)_j}{N-1} \quad (8)$$

where \mathcal{I} denotes the set of interaction pairs of interest and $R(\Gamma_i)$ denotes the rank of each interaction between feature i and the other features in input \mathbf{x} (the lowest interaction is ranked 0, and the highest interaction is ranked $N-1$). We aggregate these scores over an evaluation set to obtain a general performance score of the FIDAM. A graphical overview of this procedure is provided in Figure 2.

Baselines We consider a range of baselines in our experiments, based on the procedures explained

	Row Ranks					Relative Rank	
[4	3	0	5	2	1	5/5
(2	3	5	4	1	0	5/5
)	1	5	4	3	2	0	5/5
]	5	4	2	3	1	0	5/5
(3	2	0	1	4	5	5/5
)	3	2	0	1	5	4	5/5

ARR: 1.0

Figure 2: Example for the computation of the Average Relative Rank metric. For each row we compute the relative rank of the interaction of interest (here the Dyck language), and these row-wise relative ranks are averaged into a single score between 0 and 1. A random interaction matrix results in an ARR of around 0.5.

in §3.1. For the static baselines we consider a zero-valued baseline ($\mathbf{x}' = 0$), and a baseline that utilises a fixed mapping T based on the original input symbols ($\mathbf{x}' = T(\mathbf{x})$). Expected attributions are marginalised over samples from the distribution of well-formed strings D^+ and corrupted strings D^- . The interventional conditional expectation (Eq. 3) is computed with a corpus-wide unigram distribution ($P(x_i)$), a unigram distribution that is conditioned on the sentence position ($P(x_i|i)$), and as a joint distribution over the missing features ($P(\mathbf{x}'_S)$), that we sample from the training corpus. The observational conditional expectation (Eq. 2) is computed based on the original corpus data.²

6 Experiments on Formal Languages

We apply the evaluation procedure of §5 to two formal languages: the Identity Rule language and the Dyck-2 language. In the appendix (§A) we also present results on a palindrome language.

6.1 Identity Rule

The first language we consider is a regular language consisting of strings in which the first two symbols are identical, followed by a random sequence of symbols. The language is formed by the following grammar:

$$\begin{aligned} S &\rightarrow x x A & x &\in \{a, b, c\} \\ A &\rightarrow x A \mid \epsilon & x &\in \{a, b, c\} \end{aligned}$$

²Due to the small scale of the PCFGs considered here we can generate the complete language up to a certain length, and sample from strings that have feature overlap with the features that are still present in the partial input. For more complex tasks an auxiliary LM can be used instead.

	NB	0	$\mathbf{x}' \sim D^+$	$\mathbf{x}' \sim D^-$
Group Ablation	–	0.49	1.00	0.53
Archipelago	–	0.30	0.24	1.00
SII	–	0.70	1.00	1.00
STII	–	0.83	1.00	1.00
Hessian	0.93	–	–	–
Hessian×Input	0.66	–	–	–
IH	–	0.81	1.00	0.31

Table 1: Average Relative Rank for the Identity Rule language, columns indicate different baseline procedures. An average rank of 1 indicates that the method (correctly) assigned the interaction between the first two tokens the highest score. NB indicates these methods use no baseline.

The only interaction of interest here is between the first two symbols; all subsequent symbols are irrelevant for the prediction. An ARR score of 1.0 then indicates that for all corpus items the interaction between the first two items was the strongest out of all interactions.

We use a corpus size of 1,000, a maximum sequence length of 20, with 3 different input symbols. Corrupted strings are derived by altering one of the first two symbols (e.g. *aabcb* → *cabcb*).

Results The results for an LSTM that was trained on the language are shown in Table 1. Due to the simplicity of the language and for brevity we only report results on three baselines. A static zero-valued baseline provides imperfect interactions for all methods. The Hessian, that does not depend on any baseline, performs better than all other methods here. When sampling the baseline, however, multiple methods perfectly retrieve the interaction between the first two symbols for all corpus items. Interestingly, Group Ablation and IH benefit from sampling from the distribution of well-formed items, whereas Archipelago performs best when sampling from the distribution of corrupted items.

6.2 Dyck-2

The Dyck language is the language of well-nested brackets, and is a popular testbed for research on formal languages. It is a context-free language with center embedding clauses, requiring a model to keep track of a memory stack while processing a string. Earlier work on Dyck languages has shown that a wide range of neural model architectures can learn the grammar, including LSTMs (Sennhauser and Berwick, 2018), memory augmented RNNs (Suzgun et al., 2019), Transform-

	No baseline	Static		Expected		Interventional			Observational
		0	$T(\mathbf{x})$	D^+	D^-	$P(x'_i)$	$P(x'_i i)$	$P(\mathbf{x}'_{\setminus S})$	$P(\mathbf{x}'_{\setminus S} \mathbf{x}_S)$
Group Ablation	–	0.684	1.000	0.916	0.884	0.822	0.821	0.938	0.956
Archipelago	–	0.466	0.528	0.250	0.554	–	–	–	–
SII	–	0.555	1.000	0.921	0.895	0.876	0.885	0.923	0.989
STII	–	0.583	0.999	0.876	0.820	0.881	0.906	0.952	0.991
Hessian	0.413	–	–	–	–	–	–	–	–
Hessian×Input	0.542	–	–	–	–	–	–	–	–
IH	–	0.591	0.837	0.723	0.665	–	–	–	–

Table 2: Average Relative Ranks for the Dyck language (higher indicates stronger alignment with Dyck rules), columns indicate different baseline procedures.

ers (Ebrahimi et al., 2020), and handcrafted RNNs (Hewitt et al., 2020; Hao, 2020). We consider the Dyck-2 language, consisting of two types of brackets. The language is formed by the following grammar:

$$S \rightarrow [S] \mid (S) \mid S S \mid \epsilon$$

We use a corpus size of 15,000, a maximum sequence length of 20, and a maximum branching depth of 4. We use the same branching probabilities as Suzgun et al. (2019), which results in a uniform probability of 0.25 for each rule. Corrupted strings are derived by flipping a single bracket to any other bracket. For the baseline mapping $T(\mathbf{x})$, we map a bracket to the other bracket type, i.e. ‘(’ \leftrightarrow ‘[’ and ‘)’ \leftrightarrow ‘]’. This results in a baseline that is of the same structure as the original input, but without feature overlap.

Results We report the results for this language in Table 2, computed over all our baselines for an LSTM. The zero-valued baseline again turns out to be a mediocre baseline: for none of the methods this results in a high ARR score. The method that performs best is the fixed mapping $T(\mathbf{x})$. For Group Ablation, SII, and STII this results in a perfect ARR; for IH it is the best performing baseline.

It is encouraging that a baseline exists that results in perfect ARR scores, but this mapping depends strongly on the nature of the Dyck task itself. It is, for example, unclear how this static mapping would transfer to the natural language domain. Ideally, a more general solution makes no strong assumptions about the baseline itself. The three other baseline types in Table 2 may provide such a solution, as these only depend on the access to the original training data. Out of these, the observational baseline performs best: for the SII and STII methods this baseline performs nearly on par with the static mapping. Obtaining this conditional distribution is challenging for more complex tasks, and it can

be seen here that the interventional baseline with a joint distribution over the missing features performs well too.

7 A Natural Language Case Study: CoLA

As a case study on a larger scale natural language task, we apply our methodology to language models fine-tuned on the CoLA task (Warstadt et al., 2019). CoLA is part of the GLUE Benchmark (Wang et al., 2019), and is defined as a binary classification task of determining the linguistic acceptability of a single input sentence. The task consists of linguistically valid sentences, and sentences that contain either a syntactic, semantic, or morphological violation. A model that performs well on this task must have a thorough grasp of grammatical structure, and as such it provides a useful test bed for our FIDAM evaluation procedure.

In the previous experiments there was a degree of certainty about the structure that must be encoded by the model. In the natural language domain, however, we do not have such certainty, and should therefore be careful of making strong claims about faithfulness. Furthermore, natural language is highly multi-faceted and can not be captured by a single hierarchical structure that covers all these facets. Nonetheless, we consider it valuable to test our setup on a natural domain in order to see if interesting differences between FIDAMs arise, and whether particular facets of language such as syntactic dependency structure can be extracted.

7.1 Experimental Setup

For our experiment we consider the RoBERTa-base model (Liu et al., 2019) which obtains a Matthew’s Correlation Coefficient score of 69.70 on the in-domain validation split. We filter out sentences that contain words that are split into multiple subwords by the tokenizer, since this leads to issues with aligning the interactions of multiple subwords to

the dependency graph that is used for evaluation. Furthermore, we limit sentences to a max length of 14 in order to allow the STII and SII methods to be computed exactly without approximations. This resulted in a subset of around 60% of the original in-domain validation split that we will use in our experiment.

We evaluate the FIDAM scores on the dependency parse tree of the sentence, that we obtain with the parser of `spaCy` (Honnibal et al., 2020). The ARR score is computed based on the interaction of each token with its *parent* token. We omit the interaction of the token that has the ROOT node as its parent. An example of this procedure can be found in Appendix B. Do note that our evaluation procedure is one of many possibilities: we make the assumption that a token should interact strongly with its parent, but other interactions are likely to play a role within the model as well. We leave a more detailed investigation into using different types of linguistic structure open for future work.

We again consider the FIDAMs of Group Ablation, STII/SII, and Integrated Hessians. We leave out Archipelago, since its procedure of assigning features to a single interaction set is not feasible with our setup in which multiple child tokens might be interacting with the same parent token. Due to computational constraints we were unable to compute the full Hessian matrix of the language model, whose computation scales quadratically in the number of input *neurons* (Bishop, 2007, §5.4). For the static baselines we again consider the zero-valued baseline, as well as the `<pad>` token. The interventional baselines are obtained by computing simple count-based distributions over a sample of 100,000 sentences from the Google Books corpus. The distributions are based on the tokenization of the model’s tokenizer, and allow for computationally efficient sampling. We leave the incorporation of an observational baseline for future work, where an auxiliary masked LM might provide a useful conditional probability distribution.

7.2 Results

The results for the experiment are shown in Table 3. As expected, due to reasons outlined at the start of this section, none of the methods reaches ARR scores that are close to 1. Nonetheless, it is encouraging to see that various method/baseline combinations attain ARR scores that are far above chance level, indicating that there exists a strong

	<i>Static</i>		<i>Interventional</i>	
	0	<code><pad></code>	$P(x'_i)$	$P(\mathbf{x}'_S)$
Group Ablation	0.702	0.757	0.518	0.491
SII	0.746	0.668	0.714	0.696
STII	0.741	0.708	0.704	0.658
IH	0.577	0.516	–	–

Table 3: Average Relative Ranks for the dependency tree recovery of RoBERTa fine-tuned on CoLA.

degree of alignment between feature interactions and dependency structure. Contrary to the Dyck results, using a zero-valued baseline yields some of the highest ARR scores, which indicates that within RoBERTa’s embedding space this baseline represents a better neutral value.

A closer inspection of these results shows that the ARR scores are strongly negatively correlated to sentence length: for Group Ablation with a `<pad>` baseline, for example, we obtain a Spearman correlation of -0.38 ($p \ll 0.001$, regression plot in Appendix C). This is not surprising: as the sentence length increases, the chance of a token’s largest interaction being with its parent decreases. Another correlation of interest is between the ARR score and the model’s prediction of a sentence’s acceptability. A high correlation would indicate that the FIDAM’s alignment with dependency structure are indicative of a model’s performance. For this we obtain a Spearman correlation of 0.14 ($p = 0.036$): a relatively weak result that indicates that the structure our FIDAM extracted is only partly driving the model’s comprehension of the sentence structure.

8 Discussion & Conclusions

In this paper, we have presented a framework for characterising FIDAMs and evaluating their faithfulness. For the characterisation we set out two dimensions, feature removal and feature influence, along which existing FIDAMs can be characterised, by extending the ‘Explaining by Removing’ framework of Covert et al. to also apply to FIDAMs. This allows us to place each of the known FIDAMs in a two-dimensional grid, and to define novel variants of these models. As such, many of the methods that we incorporated in our experiments are novel FIDAMs, such as combining Archipelago with expected explanations and STII with an observational baseline.

To assess the faithfulness of FIDAMs, we made use of formal language theory and ‘grey box mod-

els’. We use formal grammars to generate multiple datasets, each with known feature interactions, and train deep learning models to perfection on those datasets. Using FIDAMs, we can then extract the learned feature interactions based on the model itself, and compare these interactions to the dependencies in the original grammar. We demonstrate that only specific combinations of FIDAMs and baselines are able to retrieve the correct interactions, while methods such as Archipelago and Integrated Hessians consistently fail to do so.

Finally, we tested our methodology on a natural language case study using a model fine-tuned on the CoLA task for linguistic acceptability. Our results on the formal language tasks either did not turn out to be predictive of this experiment or, alternatively, the results *were* predictive but the LMs made less use of dependency graph information than we might have expected. This illustrates the challenge of the Attribution Generalisation problem, and the open question remains how we can transfer faithfulness guarantees from a synthetic, controlled context to the domain of natural language and LLMs.

We do show, however, that under certain configurations feature interactions align to some degree with the (syntactic) dependency structure of a sentence. This paves the way for revealing linguistic structure in a more direct way than, for instance, can be achieved with Structural Probes (Hewitt and Manning, 2019). Investigating whether different methods and baseline configurations are able to retrieve different aspects of structure is an exciting next step that we look forward to exploring in more detail. This could be examined, for instance, through the lens of contrastive explanations Yin and Neubig (2022), a procedure that demonstrates that different baselines can reveal different aspects of linguistic structure. Furthermore, investigating the role that attention plays in modelling interactions could be a fruitful line of work, for instance by incorporating *context mixing* methods to our pipeline, such as *Value Zeroing* (Mohebbi et al., 2023) and *ALTI* (Ferrando et al., 2022).

9 Limitations

Our work has only considered *pairwise* interactions, but linguistic structure can also manifest through higher-order interactions. We show that our results on small-scale, formal languages, are different from our results on a natural language task.

It would be premature to conclude that small-scale, synthetic tasks can not be predictive of behaviour on more complex tasks, and a more detailed investigation into the properties of the task that play a role is a viable next step. Some of the FIDAMs we considered, most notably SII and STII, are intractable for larger inputs (scaling $O(2^n)$), and a necessary step in employing these methods to larger models is to construct better approximation procedures, e.g. by adapting SHAP to SII as has been done before for tabular data by Lundberg et al. (2018). More generally, although we believe our probabilistic formal language setup provides an important step forward, solving the Attribution Generalization problem – i.e., showing that results for small setups generalize to very large model – remains a key open problem.

References

- Kjersti Aas, Martin Jullum, and Anders Løland. 2021. [Explaining individual predictions when features are dependent: More accurate approximations to shapley values](#). *Artif. Intell.*, 298:103502.
- Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. 2022. [Post hoc explanations may be ineffective for detecting unknown spurious correlation](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus H. Gross. 2019. [Gradient-based attribution methods](#). In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 169–191. Springer.
- 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 2020, Online, November 16-20, 2020*, pages 3256–3274. Association for Computational Linguistics.
- Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. [“will you find these shortcuts?” a protocol for evaluating the faithfulness of input salience methods for text classification](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 976–991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Christopher M. Bishop. 2007. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1 edition. Springer.
- Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020a. [Generating hierarchical explanations on text classification via feature interaction detection](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5578–5593. Association for Computational Linguistics.
- Hugh Chen, Joseph D. Janizek, Scott M. Lundberg, and Su-In Lee. 2020b. [True to the model or true to the data?](#) *CoRR*, abs/2006.16234.
- Ian Covert, Scott M. Lundberg, and Su-In Lee. 2021. [Explaining by removing: A unified framework for model explanation](#). *J. Mach. Learn. Res.*, 22:209:1–209:90.
- Anupam Datta, Shayak Sen, and Yair Zick. 2016. [Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems](#). In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pages 598–617. IEEE Computer Society.
- 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, ACL 2020, Online, July 5-10, 2020*, pages 4443–4458. Association for Computational Linguistics.
- Javid Ebrahimi, Dhruv Gelda, and Wei Zhang. 2020. [How can self-attention networks recognize Dyck-n languages?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4301–4306, Online. Association for Computational Linguistics.
- Gabriel G. Erion, Joseph D. Janizek, Pascal Sturmfels, Scott M. Lundberg, and Su-In Lee. 2021. [Improving performance of deep learning models with axiomatic attribution priors and expected gradients](#). *Nat. Mach. Intell.*, 3(7):620–631.
- Javier Ferrando, Gerard I. Gállego, and Marta R. Costajussà. 2022. [Measuring the mixing of contextual information in the transformer](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jerome H. Friedman and Bogdan E. Popescu. 2008. Predictive learning via rule ensembles. *The annals of applied statistics*, 2(3):916–954.
- Christopher Frye, Colin Rowat, and Ilya Feige. 2020. [Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. 2016. *Deep Learning*. Adaptive computation and machine learning. MIT Press.
- Michel Grabisch and Marc Roubens. 1999. [An axiomatic approach to the concept of interaction among players in cooperative games](#). *Int. J. Game Theory*, 28(4):547–565.
- Yiding Hao. 2020. [Evaluating attribution methods using white-box LSTMs](#). In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 300–313, Online. Association for Computational Linguistics.
- John Hewitt, Michael Hahn, Surya Ganguli, Percy Liang, and Christopher D. Manning. 2020. [RNNs can generate bounded hierarchical languages with optimal memory](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1978–2010, Online. Association for Computational Linguistics.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4129–4138. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural Computation*, 9(8):1735–1780.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. [A benchmark for interpretability methods in deep neural networks](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9734–9745.
- Alon Jacovi and Yoav Goldberg. 2021. [Aligning faithful interpretations with their social attribution](#). *Trans. Assoc. Comput. Linguistics*, 9:294–310.
- Sarthak Jain, Sarah Wiegrefe, Yuval Pinter, and Byron C. Wallace. 2020. [Learning to faithfully rationalize by construction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

- Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. 2021. [Explaining explanations: Axiomatic feature interactions for deep networks](#). *J. Mach. Learn. Res.*, 22:104:1–104:54.
- Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. 2020. [Feature relevance quantification in explainable AI: A causal problem](#). In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2907–2916. PMLR.
- Yifan Jiang and Shane Steinert-Threlkeld. 2023. [The weighted möbius score: A unified framework for feature attribution](#).
- Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. [Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- 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.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2018. [Consistent individualized feature attribution for tree ensembles](#). *CoRR*, abs/1802.03888.
- Scott M. Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.
- Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. [Quantifying context mixing in transformers](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.
- W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. [Beyond word importance: Contextual decomposition to extract interactions from lstms](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic. 2022. [A song of \(dis\)agreement: Evaluating the evaluation of explainable artificial intelligence in natural language processing](#). *CoRR*, abs/2205.04559.
- Guillermo Owen. 1972. [Multilinear extensions of games](#). *Management Science*, 18(5):P64–P79.
- Nina Pörner, Hinrich Schütze, and Benjamin Roth. 2018. [Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 340–350. Association for Computational Linguistics.
- Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ["why should I trust you?": Explaining the predictions of any classifier](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.
- Alexis Ross, Matthew Peters, and Ana Marasovic. 2022. [Does self-rationalization improve robustness to spurious correlations?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Naomi Saphra and Adam Lopez. 2020. [LSTMs compose—and Learn—Bottom-up](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2797–2809, Online. Association for Computational Linguistics.
- Luzi Sennhauser and Robert C. Berwick. 2018. [Evaluating the ability of lstms to learn context-free grammars](#). In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 115–124. Association for Computational Linguistics.
- Lloyd S. Shapley. 1953. [A value for n-person games](#). *Contributions to the Theory of Games*, (28):307–317.
- 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, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.
- Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. 2021. [Integrated directional gradients: Feature interaction attribution for neural NLP models](#). In

- Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 865–878, Online. Association for Computational Linguistics.
- Chandan Singh, W. James Murdoch, and Bin Yu. 2019. [Hierarchical interpretations for neural network predictions](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Daria Sorokina, Rich Caruana, Mirek Riedewald, and Daniel Fink. 2008. [Detecting statistical interactions with additive groves of trees](#). In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 1000–1007. ACM.
- Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. 2020. [The shapley taylor interaction index](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9259–9268. PMLR.
- Mukund Sundararajan and Amir Najmi. 2020. [The many shapley values for model explanation](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278. PMLR.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. [Axiomatic attribution for deep networks](#). In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.
- Mirac Suzgun, Sebastian Gehrmann, Yonatan Belinkov, and Stuart M. Shieber. 2019. [Memory-augmented recurrent neural networks can learn generalized dyck languages](#). *CoRR*, abs/1911.03329.
- Michael Tsang, Dehua Cheng, and Yan Liu. 2018a. [Detecting statistical interactions from neural network weights](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Michael Tsang, Sirisha Rambhatla, and Yan Liu. 2020. [How does this interaction affect me? interpretable attribution for feature interactions](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Michael Tsang, Youbang Sun, Dongxu Ren, and Yan Liu. 2018b. [Can I trust you more? model-agnostic hierarchical explanations](#). *CoRR*, abs/1812.04801.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. [Neural network acceptability judgments](#). *Trans. Assoc. Comput. Linguistics*, 7:625–641.
- Sarah Wiegrefe, Ana Marasović, and Noah A. Smith. 2021. [Measuring association between labels and free-text rationales](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kayo Yin and Graham Neubig. 2022. [Interpreting language models with contrastive explanations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Matthew D. Zeiler and Rob Fergus. 2014. [Visualizing and understanding convolutional networks](#). In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer.

A Palindromes

One additional language we investigated is the context-free language of palindromes. In order to process a palindrome, a model needs to keep track of the dependency between each token in the first half of the string with its counterpart in the second half. Palindromes can contain a special symbol in the middle of a string to demarcate the two string halves, making it less ambiguous for the model at which point it should track whether the palindrome is well-formed. In our experiments, however, we found our models to perform well on both forms of palindromes. Furthermore, following [Suzgun et al. \(2019\)](#), we use a homomorphic mapping h for the second half of the string, allowing the model to use separate embeddings for symbols occurring in the first and second half of a string:

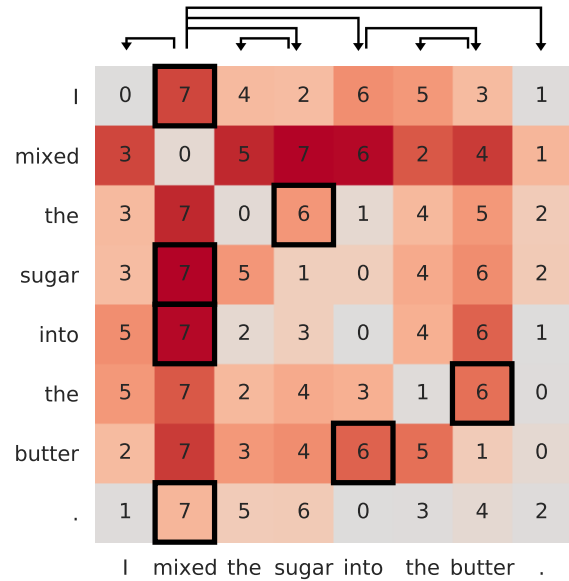
$$S \rightarrow x S h(x) \mid \epsilon \quad x \in \{a, b, c, \dots\}$$

We use a corpus size of 5,000, 10 different input symbols, and a maximum sequence length of 18. For the fixed baseline mapping $T(x)$ we map a symbol onto another random symbol, preserving the grammaticality of the palindrome (e.g. $abBA \rightarrow cdDC$).

Results The results for this language, trained with an LSTM, are shown in Figure 4. Again, the zero-valued baseline performs poorly, with most methods scoring ARR values even below chance level. The fixed baseline mapping again performs well for Group Ablation, SII, and STII, although it is not the best performing baseline this time. These three FIDAMs obtain perfect performance when using the expected baselines over a distribution of well-formed palindromes, which also holds for the interventional baseline with a joint distribution over the missing features. This is in contrast to the Dyck results, where the observational baseline resulted in better ARR scores for all three of these methods.

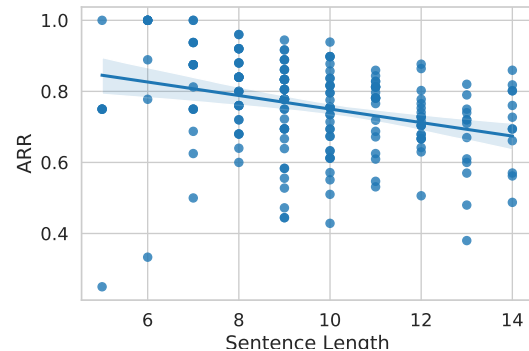
B ARR Example

An example of a sentence with a high ARR (0.93), for the Group Ablation method with a `<pad>` baseline:



C Correlation CoLA ARR and sentence length

Correlation between sentence length and ARR, shown here for Group Ablation with a `<pad>` baseline. Spearman's $\rho = -0.38$ ($p \ll 0.001$):



	<i>Static</i>		<i>Expected</i>		<i>Interventional</i>			<i>Observational</i>
	0	$T(\mathbf{x})$	D^+	D^-	$P(x'_i)$	$P(x'_i i)$	$P(\mathbf{x}'_{\setminus S})$	$P(\mathbf{x}'_{\setminus S} \mathbf{x}_S)$
Group Ablation	0.450	0.980	1.000	0.943	0.777	0.836	1.000	0.939
Archipelago	0.356	0.452	0.439	0.717	–	–	–	–
SII	0.472	0.933	1.000	0.892	0.804	0.817	1.000	1.000
STII	0.472	0.921	0.999	0.917	0.760	0.792	1.000	0.999
Hessian	0.523	–	–	–	–	–	–	–
Hessian×Input	0.523	–	–	–	–	–	–	–
IH	0.505	0.637	0.693	0.535	–	–	–	–

Table 4: Average Relative Ranks for the palindrome language (higher is better).

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Section 9
- A2. Did you discuss any potential risks of your work?
Our work is of a more theoretical nature, providing a new way of measuring the faithfulness of feature interaction methods.
- A3. Do the abstract and introduction summarize the paper’s main claims?
Section 1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Left blank.

- B1. Did you cite the creators of artifacts you used?
No response.
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C Did you run computational experiments?

Section 6 & 7

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We report model details. The small scale of our setup allows it to be ran on any normal contemporary laptop, and as such we do not report GPU hourse.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Section 6

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

No response.

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Mentioned in Section 7 (spaCy).

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.