

AttnLRP: Attention-Aware Layer-wise Relevance Propagation for Transformers

Reduan Achtabat¹ Sayed Mohammad Vakilzadeh Hatefi¹ Maximilian Dreyer¹
 Aakriti Jain¹ Thomas Wiegand^{1,2,3} Sebastian Lapuschkin^{1,†} Wojciech Samek^{1,2,3,†}

¹ Fraunhofer Heinrich-Hertz-Institute, 10587 Berlin, Germany

² Technische Universität Berlin, 10587 Berlin, Germany

³ BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany

[†] corresponding authors: {wojciech.samek, sebastian.lapuschkin}@hhi.fraunhofer.de

Abstract

Large Language Models are prone to biased predictions and hallucinations, underlining the paramount importance of understanding their model-internal reasoning process. However, achieving faithful attributions for the entirety of a black-box transformer model and maintaining computational efficiency is an unsolved challenge. By extending the Layer-wise Relevance Propagation attribution method to handle attention layers, we address these challenges effectively. While partial solutions exist, our method is the first to faithfully and holistically attribute not only input but also latent representations of transformer models with the computational efficiency similar to a singular backward pass. Through extensive evaluations against existing methods on Llama 2, Flan-T5 and the Vision Transformer architecture, we demonstrate that our proposed approach surpasses alternative methods in terms of faithfulness and enables the understanding of latent representations, opening up the door for concept-based explanations. We provide an open-source implementation on GitHub¹.

1. Introduction

The attention mechanism (Vaswani et al., 2017) became an essential component of large transformers due to its unique ability to handle multimodality and to scale to billions of training samples. While these models demonstrate impressive performance in text and image generation, they are prone to biased predictions and hallucinations (Huang et al., 2023), which hamper their widespread adoption.

To overcome these limitations, it is crucial to understand

¹<https://github.com/rachtibat/LRP-for-Transformers>

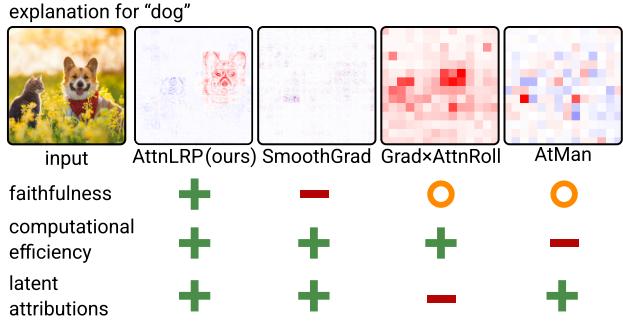


Figure 1. By optimizing LRP for transformer-based architectures, our LRP variant outperforms other state-of-the-art methods in terms of explanation faithfulness and computational efficiency. We further are able to explain latent neurons inside and outside the attention module, allowing us to interact with the model. A more detailed discussion on the differences between AttnLRP and other LRP variants can be found in Appendix A.2.2. Heatmaps for other methods are illustrated in Appendix Figure 6. Legend: highly (+), semi- (○), not suited (−). Credit: Nataba/iStock.

the latent reasoning process of transformer models. Researchers started using the attention mechanism of transformers as a means to understand how input tokens interact with each other. Attention maps contain rich information about the data distribution (Clark et al., 2019; Caron et al., 2021), even allowing for image data segmentation. However, attention, by itself, is inadequate for comprehending the full spectrum of model behavior (Wiegreffe and Pinter, 2019). Similar to latent activations, attention is not class-specific and solely provides an explanation for the softmax output (in attention layers) while disregarding other model components. Recent works (Geva et al., 2021; Dai et al., 2022), e.g., have discovered that factual knowledge in Large Language Models (LLMs) is stored in Feed-Forward Network (FFN) neurons, separate from attention layers. Further, attention-based attribution methods such as rollout (Abnar and Zuidema, 2020; Chefer et al., 2021a)

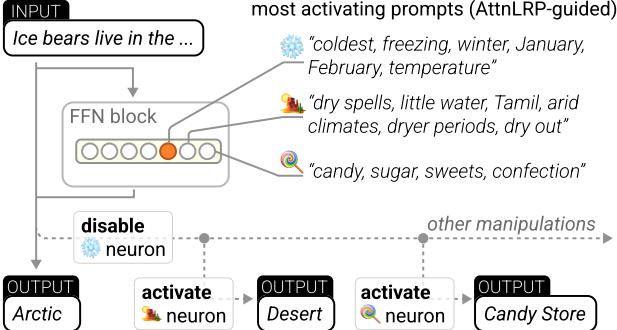


Figure 2. AttnLRP combined with ActMax allows to identify relevant neurons and gain insights into their encodings. This allows one to manipulate the latent representations and, *e.g.*, to change the output “Arctic” (by disabling the corresponding neuron) to “Desert” or “Candy Store” (by activating the respective neurons). See also Section 4.3.

result in checkerboard artifacts, as visible in Figure 1 for a Vision Transformer (ViT). Researchers thus have turned to model-agnostic approaches that aim to provide a holistic explanation of the model’s behavior (Miglani et al., 2023), including, *e.g.*, perturbation and gradient-based methods. Methods based on feature perturbation require excessive amounts of compute time (and energy), making their application uneconomically, especially for large architectures. Gradient-based methods, on the other hand, are highly efficient but suffer from noisy gradients and low faithfulness, as evaluated in Section 4.1. Consequently, for understanding latent neurons, perturbation methods remain often the preferred choice (Meng et al., 2022; Bills et al., 2023).

Another option is to take advantage of the versatility of rule-based backpropagation methods, such as Layer-wise Relevance Propagation (LRP). These methods allow for the customization of propagation rules to accommodate novel operations, allowing for more faithful explanations and requiring only a singular backward pass. Attempts to apply LRP to non-linear attention, however, have faced challenges such as numerical instability, lack of faithfulness, or completely circumventing the non-linear softmax problem, see Appendix A.2.2.

Contributions In this work, we introduce AttnLRP, an extension of LRP within the Deep Taylor Decomposition framework (Montavon et al., 2017), with the particular requirements necessary for attributing non-linear attention accurately. AttnLRP allows explaining transformer-based models with high faithfulness and efficiency, while also allowing attribution of latent neurons and providing insights into their role in the generation process (see Figure 2).

1. We derive novel efficient and faithful LRP attribution rules for non-linear attention within the Deep Taylor

Decomposition framework, demonstrating their superiority over the state-of-the-art and successfully tackling the noise problem in ViTs.

2. We illustrate how to gain insights into a LLM generation process by identifying relevant neurons and explaining their encodings.
3. We provide an efficient and ready-to-use open source implementation of AttnLRP for transformers.

2. Related Work

We present an overview of related work for various model-agnostic and transformer-specialized attribution methods.

2.1. Perturbation & Local Surrogates

In perturbation analysis, such as occlusion-based attribution (Zeiler and Fergus, 2014) or SHAP (Lundberg and Lee, 2017), the input features are repeatedly perturbed while the effect on the model output is measured (Fong and Vedaldi, 2017). AtMan (Deb et al., 2023) is specifically adapted to the transformer architecture, where tokens are not suppressed in the input space, but rather in the latent attention weights. Interpretable local surrogates, on the other hand, replace complex black-box models with simpler linear models that locally approximate the model function being explained. Since the surrogate has low complexity, interpretability is facilitated. Prominent methods include LIME (Ribeiro et al., 2016) and LORE (Guidotti et al., 2018).

While these approaches are model-agnostic and memory efficient, they have a high computational cost in terms of forward passes. Furthermore, explanations generated on surrogate models cannot explain the hidden representations of the original model. Finally, latent attributions wrt. the prediction must be computed for each layer separately, increasing the computational cost further.

2.2. Attention-based

These methods take advantage of the attention mechanism in transformer models. Although attention maps capture parts of the data distribution, they lack class specificity and do not provide a meaningful interpretation of the final prediction (Wiegrefe and Pinter, 2019). Attention Rollout (Abnar and Zuidema, 2020) attempts to address the issue by sequentially connecting attention maps of all layers. However, the resulting attributions are still not specific to individual outputs and exhibit substantial noise. Hence, (Gildenblat, 2023) has found that reducing noise in attention rollout can be achieved by filtering out excessively strong outlier activations. To enable class-specificity, the work of (Chefer et al., 2021b) proposed a novel rollout pro-

cedure wherein the attention’s activation is mean-weighted using a combination of the gradient and LRP-inspired relevances. It is important to note that this approach yields an approximation of the mean squared relevance value, which diverges from the originally defined notion of “relevance” or “importance” of additive explanatory models such as SHAP (Lundberg and Lee, 2017) or LRP (Bach et al., 2015). Subsequent empirical observations by (Chefer et al., 2021a) revealed that an omission of LRP-inspired relevances and a sole reliance on a positive mean-weighting of the attention’s activation with the gradient improved the faithfulness inside cross-attention layers. Though, this approach can only attribute positively and does not consider counteracting evidence.

Attention-rollout based approaches, while offering advantages in terms of computational efficiency and conceptual simplicity, have significant drawbacks. Primarily, they suffer from a limited resolution in the input attribution maps, resulting in undesirable checkerboard artifacts cf. Figure 1. Moreover, they are unable to attribute hidden latent features beyond the softmax output. Consequently, these approaches only provide explanations for a fraction of the model, thereby compromising the overall fidelity and limiting the feasibility of explanations within the hidden space.

2.3. Backpropagation-based

Input \times Gradient (Simonyan et al., 2014) linearizes the model by utilizing the gradient. However, this approach is vulnerable to gradient shattering (Balduzzi et al., 2017; Dombrowski et al., 2022), leading to noisy attributions in deep models. Consequently, several works aim to denoise these attributions. SmoothGrad (Smilkov et al., 2017) and Integrated Gradients (Sundararajan et al., 2017) have attempted to address this issue but have been unsuccessful in the case of large transformers, as demonstrated in the experiments in Section 4.1. (Chefer et al., 2021b) adapted Grad-CAM (Selvaraju et al., 2017) to transformer models by weighting the last attention map with the gradient.

Modified backpropagation methods, such as LRP (Bach et al., 2015), decompose individual layers instead of linearizing the entire model. They modify the gradient to produce more reliable attributions (Arras et al., 2022). The work (Ding et al., 2017) was the first to apply standard LRP on non-linear attention layers, while (Voita et al., 2021) proposed an improved variant building upon Deep Taylor decomposition. Nonetheless, both variants can lead to numerical instabilities in the softmax and do not fulfill the conservation property (3) in matrix multiplication (Lemma 3 of (Chefer et al., 2021b)). (Ali et al., 2022) considerably improved attributions by regarding the non-linear softmax operation as constant attributing relevance solely through the value path and disregarding the softmax operation en-

tirely. In Appendix A.2.2, we compare the different LRP-variants.

3. Attention-Aware LRP for Transformers

First, we motivate LRP in the framework of additive explanatory models. Then, we generalize the design of new rules for non-linear operations. Finally, we apply our methodology successively on each operation utilized in a transformer model to derive efficient and faithful rules.

3.1. Layer-wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) (Bach et al., 2015; Montavon et al., 2019) belongs to the family of additive explanatory models, which includes the well-known Shapley (Lundberg and Lee, 2017), Gradient \times Input (Simonyan et al., 2014) and DeepLIFT (Shrikumar et al., 2017) methods.

The underlying assumption of such models is that a function f_j with N input features $\mathbf{x} = \{x_i\}_{i=1}^N$ can be decomposed into individual contributions of single input variables $R_{i \leftarrow j}$ (called “relevances”). Here, $R_{i \leftarrow j}$ denotes the amount of output j that is attributable to input i , which, when added together, equals (or is proportional to) the original function value. Mathematically, this can be written as:

$$f_j(\mathbf{x}) \propto R_j = \sum_i^N R_{i \leftarrow j} \quad (1)$$

If an input i is connected to several outputs j , e.g., a multidimensional function \mathbf{f} , the contributions of each output j are losslessly aggregating together.

$$R_i = \sum_j R_{i \leftarrow j}. \quad (2)$$

This provides us with “importance values” for the input variables, which reveal their direct contribution to the final prediction.

Unlike other methods, LRP treats a neural network as a layered directed acyclic graph, where each neuron j in layer l is modeled as a function node f_j^l that is individually decomposed according to equation (1). Beginning at the model output L , the initial relevance value $R_j^L \propto f_j^L$ is successively distributed to its prior network neurons one layer at a time. Hence, LRP follows the flow of activations computed during the forward pass through the model in the *opposite* direction, from output f^L back to input layer f^1 .

This decomposition characteristic of LRP gives rise to the important *conservation property*:

$$R^{l-1} = \sum_i R_i^{l-1} = \sum_i \sum_j R_{i \leftarrow j}^{l-1} = \sum_j R_j^l = R^l, \quad (3)$$

ensuring that the sum of all relevance values in each layer remains constant. This property allows for meaningful attribution, as the scale of each relevance value can be related to the original function output f^L .

3.1.1. DECOMPOSITION THROUGH LINEARIZATION

To design a faithful attribution method, the challenge lies in identifying a meaningful distribution rule $R_{i \leftarrow j}$. Possible solutions encompass all decompositions that adhere to the conservation property (3). However, for a decomposition to be considered *faithful*, it should approximate the characteristics of the original function as closely as possible.

In this paper, we take advantage of the Deep Taylor Decomposition framework (Montavon et al., 2017) to locally linearize and decompose neural network operations into independent contributions. As a special case, we further establish the relationship between one derived rule and the Shapley Values framework in Section 3.3.2.

We start by computing a first-order Taylor expansion at a reference point $\tilde{\mathbf{x}}$. For the purpose of simplifying the equation, we assume that the reference point $\tilde{\mathbf{x}}$ is constant:

$$\begin{aligned} f_j(\mathbf{x}) &= f_j(\tilde{\mathbf{x}}) + \sum_i \mathbf{J}_{ji}(\tilde{\mathbf{x}}) (x_i - \tilde{x}_i) + \mathcal{O}(|\mathbf{x} - \tilde{\mathbf{x}}|^2) \quad (4) \\ &= \sum_i \mathbf{J}_{ji} x_i + f_j(\tilde{\mathbf{x}}) - \underbrace{\sum_i \mathbf{J}_{ji} \tilde{x}_i}_{\text{bias } \tilde{b}_j} + \mathcal{O}(|\mathbf{x} - \tilde{\mathbf{x}}|^2) \end{aligned}$$

where \mathcal{O} is the approximation error in Big- \mathcal{O} notation and the Jacobian \mathbf{J} is evaluated at reference point $\tilde{\mathbf{x}}$, that is in the following omitted for brevity². The bias term represents the constant portion of the function and the approximation error that cannot be directly attributed to the input variables.

We substitute the neural network function with its first-order expansion and assert its proportionality to a relevance value R_j through multiplication with a constant factor $c \in \mathbb{R}$ with $f_j(\mathbf{x}) \neq 0$.

$$R_j = f_j(\mathbf{x}) c = \left(\sum_i \mathbf{J}_{ji} x_i + \tilde{b}_j \right) \frac{R_j}{f_j(\mathbf{x})}$$

Generally, we focus on attributing the input variables and hence ignore the contribution of the bias term (Bach et al., 2015). However, it is important to note that the bias absorbs part of the relevance and that the conservation property (3) still holds if the bias is regarded as an additional neuron without connection to the input. Alternatively, the bias term can be distributed equally among the input variables, as explained in Appendix A.2.1.

Comparing with Equation (1), we identify $R_{i \leftarrow j}$ as the summands and apply Equation (2) to the input variables. In

addition, we insert stabilizing factor ε (usually set to 10^{-6}) with the sign of $f_j(\mathbf{x})$ to allow for the case $f_j(\mathbf{x}) = 0$:

$$R_i = \sum_j R_{i \leftarrow j} = \sum_j \mathbf{J}_{ji} x_i \frac{R_j}{f_j(\mathbf{x}) + \varepsilon \operatorname{sign}(f_j(\mathbf{x}))} \quad (5)$$

In the following, $\operatorname{sign}(f_j(\mathbf{x}))$ is omitted for brevity.

To benefit from GPU parallelization, this formula can be written in matrix form:

$$\Rightarrow \mathbf{R}^{l-1} = \mathbf{x} \odot \mathbf{J}^T \cdot \mathbf{R}^l \oslash (\mathbf{f}(\mathbf{x}) + \varepsilon)$$

where \odot denotes the Hadamard product and \oslash element-wise division. This formula can be efficiently implemented in automatic differentiation libraries, such as PyTorch (Paszke et al., 2019). Compared to a basic backward pass, we have additional computational complexity for the element-wise operations.

3.2. ATTRIBUTING THE MULTILAYER PERCEPTRON

Commonly a Multilayer Perceptron consists of a linear layer with a (component-wise) non-linearity producing input activations for the succeeding layer(s):

$$z_j = \sum_i \mathbf{W}_{ji} x_i + b_j \quad (6)$$

$$a_j = \sigma(z_j) \quad (7)$$

where \mathbf{W}_{ji} are the weight parameters and σ constitutes a (component-wise) non-linearity.

3.2.1. THE ε - AND γ -LRP RULE

Linearizing linear layers (6) at any point $\mathbf{x} \in \mathbb{R}^N$ results in the fundamental ε -LRP (Bach et al., 2015) rule

$$R_i = \sum_j \mathbf{W}_{ji} x_i \frac{R_j}{z_j(\mathbf{x}) + \varepsilon} \quad (8)$$

The bias b_j (Equation (6)) absorbs a portion of the relevance. The proof is omitted for brevity. We employ the ε -LRP rule on all linear layers, unless specified otherwise.

In models with many layers, the gradient of a linear layer (equal to the weight matrix \mathbf{W}) can cause noisy attributions due to the gradient shattering effect (Balduzzi et al., 2017; Dombrowski et al., 2022). To mitigate this noise, it is best practice to use the γ -LRP rule (Montavon et al., 2019), an extension to improve the signal-to-noise ratio. We have observed that this effect is significantly pronounced in ViTs while LLMs lack visible noise. Therefore, we only apply the γ -LRP rule to linear layers in ViTs. For more details, please refer to Appendix A.2.3.

²if $\tilde{\mathbf{x}} = \mathbf{x}$, this is equivalent to Gradient \times Input

3.2.2. HANDLING ELEMENT-WISE NON-LINEARITIES

Since element-wise non-linearities have only a single input and output variable, the decomposition of equation (1) is the operation itself. Therefore, the entire incoming relevance R_j^l can only be assigned to the single input variable.

$$R_i^{l-1} = R_i^l \quad (9)$$

The identity rule (9) is applied to all element-wise operations with a single input and single output variable.

3.3. ATTRIBUTING NON-LINEAR ATTENTION

The heart of the transformer architecture (Vaswani et al., 2017) is non-linear attention

$$\mathbf{A} = \text{softmax} \left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}} \right) \quad (10)$$

$$\mathbf{O} = \mathbf{A} \cdot \mathbf{V} \quad (11)$$

$$\text{softmax}_j(\mathbf{x}) = \frac{e^{x_j}}{\sum_k e^{x_k}} \quad (12)$$

where (\cdot) denotes matrix multiplication, $\mathbf{K} \in \mathbb{R}^{b \times k \times d_k}$ is the key matrix, $\mathbf{Q} \in \mathbb{R}^{b \times q \times d_k}$ is the queries matrix, and $\mathbf{V} \in \mathbb{R}^{b \times k \times d_v}$ the values matrix, and $\mathbf{O} \in \mathbb{R}^{b \times k \times d_v}$ is the final output of the attention mechanism. b is the batch dimension including the number of heads, and d_k, d_v indicate the embedding dimensions, and s_q, s_k are the number of query and key/value tokens.

First and foremost, the softmax function is highly non-linear. In addition, the matrix multiplication is bilinear, *i.e.*, linear in both of its input variables. In the following, we will derive relevance propagation rules for each of these operations, taking into account considerations of efficiency.

3.3.1. HANDLING THE SOFTMAX NON-LINEARITY

In Section 3.1.1, we present a generalized approach to linearization that incorporates bias terms, allowing for the absorption of a portion of the relevance. However, (Ali et al., 2022) advocates for a strict adherence to the conservation property (3) and argues that a linear decomposition of a *non-linear* function should typically exclude a bias term. While we see the virtue of this approach for operations such as RMSNorm (Zhang and Sennrich, 2019) or matrix multiplication, where $f(0) = 0$, we contend that a linearization of the softmax function should inherently incorporate a bias term. This is due to the fact that even when the input is zero, the softmax function yields a value of $\frac{1}{N}$ (where N represents the dimension of the inputs) which is analogous to a bias term.

Proposition 3.1 *Decomposing the softmax function by a Taylor decomposition (4) at reference point \mathbf{x} yields the following relevance propagation rule:*

$$R_i^l = x_i(R_i^{l+1} - s_i \sum_j R_j^{l+1}) \quad (13)$$

where s_j denotes the j -th output of the softmax function. The hidden bias term, which represents the approximation error, consequently absorbs a portion of the relevance.

The proof can be found in Appendix A.3.1.

Note, that the works (Voita et al., 2021; Chefer et al., 2021b; Ali et al., 2022) propose to handle the bias term differently to strictly enforce the conservation property (3). (Voita et al., 2021) linearizes at \mathbf{x} but distributes the bias term equally on all input variables, while (Chefer et al., 2021b) applies the element-wise identity rule (9) and hence omits the bias term completely. Both variants can lead to severe numerical instabilities as discussed in Appendix A.2.1 and seen empirically in our preliminary experiments. Finally, (Ali et al., 2022) regards the attention matrix \mathbf{A} in equation (11) as constant, attributing relevance solely through the value path and disregarding the softmax operation entirely. A more detailed discussion on the differences of the LRP variants can be found in Appendix A.2.2.

3.3.2. HANDLING MATRIX-MULTIPLICATION

Since $f(0, 0) = 0$ holds, it is desirable to decompose the matrix multiplication without a bias term. To achieve this, we break down the matrix multiplication into an affine operation involving summation and a bi-linear part involving element-wise multiplication.

$$\mathbf{O}_{jp} = \sum_i \underbrace{\mathbf{A}_{ji} \mathbf{V}_{ip}}_{\text{bi-linear part}}$$

The summation already provides a decomposition in the form of equation (1), and we only need to decompose the individual summands $\mathbf{A}_{ji} \mathbf{V}_{ip}$. Since multiplication is a commutative operation, assigning equal relevance values to both operands is reasonable.

Proposition 3.2 *Decomposing element-wise multiplication with N input variables of the form*

$$f_j(\mathbf{x}) = \prod_i^N x_i$$

by Shapley (with baseline zero) or Taylor decomposition (4) at reference point \mathbf{x} (without bias or distributing the bias uniformly) yields the following uniform relevance propagation rule:

$$R_{i \leftarrow j}(x_i) = \frac{1}{N} R_j. \quad (14)$$

The proof can be found in Appendix A.3.2. Consequently, the combined rule can be effectively computed using:

Proposition 3.3 *Decomposing matrix multiplication with a sequential application of the uniform rule (14) and the ε -rule (8) yields the following relevance propagation rule:*

$$R_{ji}^{l-1}(\mathbf{A}_{ji}) = \sum_p \mathbf{A}_{ji} \mathbf{V}_{ip} \frac{R_{jp}^l}{2 \mathbf{O}_{jp} + \varepsilon} \quad (15)$$

There is no bias term absorbing relevance. For \mathbf{V}_{ip} , we sum over the j indices. The proof can be found in Appendix A.3.3. By employing this rule, we are able to maintain strict adherence to the conservation property (3), while also preserving numerical stability.

3.3.3. HANDLING NORMALIZATION LAYERS

Commonly used normalization layers in Transformers include LayerNorm (Ba et al., 2016) and RMSNorm (Zhang and Sennrich, 2019). These layers apply affine transformations and non-linear normalization sequentially.

$$\text{LayerNorm}(\mathbf{x}) = \frac{\mathbf{x}_j - \mathbb{E}[\mathbf{x}]}{\sqrt{\text{Var}[\mathbf{x}] + \varepsilon}} \gamma_j + \beta_j \quad (16)$$

$$\text{RMSNorm}(\mathbf{x}) = \frac{\mathbf{x}_j}{\sqrt{\frac{1}{N} \sum_k \mathbf{x}_k^2 + \varepsilon}} \gamma_j \quad (17)$$

where $\varepsilon, \gamma_j, \beta_j \in \mathbb{R}$. Affine transformations such as the multiplicative weighting of the output or the subtraction of the mean value are linear operations that can be attributed by the ε -LRP rule. Normalization, on the other hand, is non-linear and requires separate considerations. As such, we focus on the following function:

$$f_j(\mathbf{x}) = \frac{\mathbf{x}_j}{g(\mathbf{x})} \quad (18)$$

where $g(\mathbf{x}) = \sqrt{\text{Var}[\mathbf{x}] + \varepsilon}$ or $g(\mathbf{x}) = \sqrt{\frac{1}{N} \sum_k \mathbf{x}_k^2 + \varepsilon}$.

The work (Ali et al., 2022) demonstrates that when linearizing LayerNorm at \mathbf{x} , the bias term absorbs most of the relevance equal to $\text{Var}[\mathbf{x}] / (\text{Var}[\mathbf{x}] + \varepsilon)$, effectively absorbing 99% of the relevance with commonly used values of $\varepsilon = 10^{-6}$ and $\text{Var}[\mathbf{x}] = 1$. Hence, a linearization at \mathbf{x} is not meaningful. As a solution, (Ali et al., 2022) proposes to skip this operation with the identity rule. In the following, we prove that this heuristic can be derived from the Deep Taylor Decomposition framework.

Proposition 3.4 *Decomposing LayerNorm or RMSNorm by a Taylor decomposition (4) with reference point $\mathbf{0}$ (without bias or distributing the bias uniformly) yields the identity relevance propagation rule:*

$$R_i^{l-1} = R_i^l \quad (19)$$

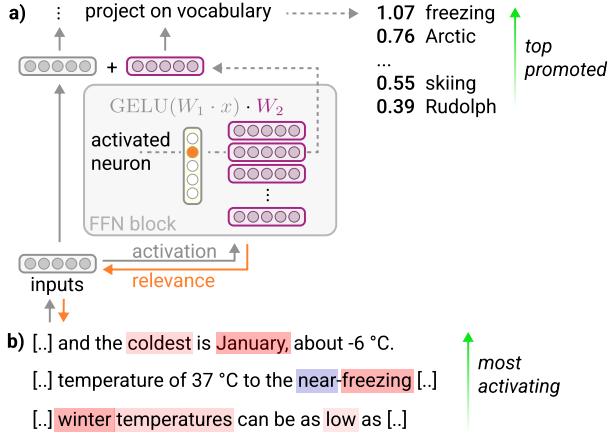


Figure 3. There are two approaches for understanding knowledge neurons: (a) Neuron 3948 at the last non-linearity in FFN 17 of the Phi-1.5 model selects a weight row to add to the residual stream. This weight row projected on the vocabulary spans topics about ice, cold places and winter sport. (b) Sentences that maximally activate this neuron contain references about coldness. Attributing the neuron with AttnLRP highlights the most relevant tokens inside the input sentences. Note, although not illustrated here, AttnLRP can also attribute linear layers inside attention modules. Inspired by (Voita et al., 2023).

There is no bias that absorbs relevance. The proof is given in Appendix A.3.4

This rule enforces a strict notion of conservation, while being highly efficient by excluding normalization operations from the computational graph. Experiments in Section 4.1 provide evidence that this simplification is faithful.

3.4. Understanding Latent Features

As we iterate through each layer and neuron during the attribution process with AttnLRP, we obtain relevance values for each latent neuron as a by-product, which directly signify their importance for the prediction. Ranking this latent relevance enables us to identify the neurons and layers that are most influential for the reasoning process of the model (Achitbat et al., 2023). After identifying a group of neurons, the subsequent step is to reveal the concept that is represented by each neuron.

For Convolutional Neural Networks (CNNs), single units (*i.e.*, neurons or filters) have been shown to correspond to distinct concepts, each fulfilling specific sub-tasks (Radford et al., 2017). In order to find the most representative reference samples that explain the neuron's encoding, researchers rely on Activation Maximization (Act-Max) (Nguyen et al., 2016), where input samples are sought that give rise to the highest activation value. We follow up on these observations and present the following strategy for understanding latent features:

(1) Collect prompts that lead to the highest activation of a unit. (2) Explain the unit’s activation using AttnLRP, allowing to narrow down the relevant input tokens for the chosen unit.

In this work, we concentrate on knowledge neurons (Dai et al., 2022; Voita et al., 2023) that are situated at the last non-linearity in FFN layers $z_j = \text{GELU}(W_1\mathbf{x})$. These neurons possess intriguing properties, as shown in Figure 3: They encode factual knowledge and upon activation, the corresponding row of the second weight matrix W_2 is added to the residual stream directly influencing the output distribution of the model. By projecting this weight row onto the vocabulary, a distribution of the most probable tokens across the vocabulary is obtained (Geva et al., 2022). Applying AttnLRP on ActMax reference samples and projecting the weight row on the vocabulary allow us to understand in which context a neuron activates and how its activation influences the prediction of the next token. In contrast to (Ali et al., 2022), AttnLRP also allows analyzing the key and value linear layers inside attention modules.

4. Experiments

Our experiments aim to answer the following questions:

- (Q1)** How faithful are our explanations compared to other state-of-the-art approaches?
- (Q2)** How efficient is LRP compared to perturbation-based methods?
- (Q3)** Can we understand latent representations and interact with LLMs?

4.1. Evaluating Explanations (Q1)

A reliable measure of faithfulness of an explanation are input perturbation experiments (Samek et al., 2017). This approach iteratively substitutes the most important tokens in the input domain with a baseline value. If the attribution method accurately identified the most important tokens, the model’s confidence in the predicted output should rapidly decrease. The other way around, perturbing the least relevant tokens first, should not affect the model’s prediction and result in a slow decline of the model’s confidence. For more details, see Appendix B.2. Despite its drawbacks, such as potentially introducing out-of-distribution manipulations (Chang et al., 2018) and sensitivity towards the chosen baseline value, this approach is widely adopted in the community. (Blücher et al., 2024) has addressed this criticism and introduced an enhanced metric by quantifying the area between the least and most relevant order perturbation curves to obtain a robust measure. Hence, we will employ this improved metric to measure faithfulness. Appendix Figure 4 illustrates a typical perturbation curve.

In order to assess plausibility, we utilize the SQuAD v2 Question-Answering (QA) dataset (Rajpurkar et al., 2018), which includes a ground truth mask indicating the correct answer within the question. We calculate attributions for accurately answered questions and determine the Intersection over Union (IoU) between the positive attribution values and the ground truth mask. This approach assumes that the model solely relies on the information provided in the ground truth mask, which is not entirely accurate but sufficient for identifying a trend.

4.1.1. BASELINES

We evaluate the faithfulness on two self-attention models, a ViT-B-16 (Dosovitskiy et al., 2021) on ImageNet (Deng et al., 2009) classification and the Llama 2-7b (Touvron et al., 2023) model on IMDB movie review (Maas et al., 2011) classification as well as next word prediction of Wikipedia (Wikimedia Foundation, 2023). To evaluate cross-attention layers, we use the encoder-decoder model Flan T5-XL (Chung et al., 2022) on the SQuAD v2 dataset. We denote our method as AttnLRP and compare it against a broad spectrum of methods including Input \times Gradient (I \times G), Integrated Gradients (IG), SmoothGrad (SmoothG), Attention Rollout (AttnRoll), Gradient-weighted Attention Rollout (G \times AttnRoll) and Conservative Propagation (CP)-LRP. As explained in Appendix A.2.3, we propose to apply the γ -rule for AttnLRP in the case of ViTs. For better comparison, we also included an enhanced CP-LRP baseline, which also uses the γ -rule in the ViTs experiment. The LRP variants introduced by (Voita et al., 2021; Chefer et al., 2021b) are excluded due to numerical instabilities observed in preliminary experiments, see also Appendix A.2.1. Further, we utilize the Grad-CAM adaptation described in (Chefer et al., 2021b). Specifically, we weight the last attention map with the gradient. Finally, we expand upon AtMan by incorporating it into encoder-decoder models by suppressing tokens in all self-attention layers within the encoder, while only doing so in cross-attention layers within the decoder. For AtMan, SmoothGrad and Rollout-methods we perform a hyperparameter sweep over a subset of the dataset. More details are in Appendix B.3.

4.1.2. DISCUSSION

In Table 1, it is visible that AttnLRP consistently outperforms all the state-of-the-art methods in terms of faithfulness, especially in Llama 2. We also observe that gradient-based approaches significantly suffer from noisy attributions, as reflected by the low faithfulness scores and illustrated in example heatmaps in Appendix B.5. CP-LRP with ε applied on all layers (as proposed in Ali et al. (2022)), also suffers from noisy gradients in ViTs. Applying the γ -rule for CP-LRP and AttnLRP in ViTs improves the faithfulness substantially. Whereas AtMan does not perform

Table 1. Faithfulness scores as area between the least and most relevant order perturbation curves (Blücher et al., 2024) on different models and datasets. To assess plausibility, the IoU in the SQuAD v2 dataset is depicted. The higher a score, the better.

Methods	ViT-B-16		Llama 2-7b		Flan-T5-XL	
	ImageNet	IMDB	Wikipedia	SQuAD v2	SQuAD v2	IoU
Random	0.01 ± 0.01	-0.01 ± 0.05	-0.07 ± 0.13	0.01 ± 0.21	0.079 ± 0.001	
Input \times Grad (Simonyan et al., 2014)	0.80 ± 0.03	0.12 ± 0.05	0.18 ± 0.13	0.27 ± 0.14	0.386 ± 0.003	
IG (Sundararajan et al., 2017)	1.54 ± 0.03	1.23 ± 0.05	4.05 ± 0.13	0.77 ± 0.14	0.158 ± 0.002	
SmoothGrad (Smilkov et al., 2017)	-0.04 ± 0.03	0.25 ± 0.05	-2.22 ± 0.14	0.16 ± 0.15	0.090 ± 0.001	
GradCAM (Chefer et al., 2021b)	0.27 ± 0.04	-0.82 ± 0.05	2.01 ± 0.15	0.94 ± 0.14	0.692 ± 0.004	
AttnRoll (Abnar and Zuidema, 2020)	1.31 ± 0.03	-0.64 ± 0.05	-3.49 ± 0.15	-0.42 ± 0.14	0.076 ± 0.001	
Grad \times AttnRoll (Chefer et al., 2021a)	2.60 ± 0.03	1.61 ± 0.05	9.79 ± 0.14	-0.06 ± 0.14	0.531 ± 0.002	
AtMan (Deb et al., 2023)	0.70 ± 0.02	-0.20 ± 0.05	3.31 ± 0.15	1.01 ± 0.15	0.798 ± 0.003	
CP-LRP (ε -rule, Ali et al. (2022))	2.53 ± 0.02	1.72 ± 0.04	7.85 ± 0.12	1.74 ± 0.14	0.830 ± 0.003	
CP-LRP (γ -rule for ViT, as proposed here)	6.06 ± 0.02	-	-	-	-	
AttnLRP (ours)	6.19 ± 0.02	2.50 ± 0.05	10.93 ± 0.13	1.76 ± 0.14	0.840 ± 0.002	

well in unstructured tasks, *i.e.*, next word prediction or classification, it achieves a high score in QA tasks. While G \times AttnRoll better reflects the model behavior in unstructured tasks compared to AtMan, it is affected by considerable background noise, resulting in a low IoU score in the SQuAD v2 dataset.

4.2. Computational Complexity and Memory Consumption (Q2)

Table 2 illustrates the computational complexity and memory consumption of a single LRP-based attribution and linear-time perturbation, such as AtMan or a Shapley-based method (Fatima et al., 2008). Linear-time perturbation requires N_T forward passes, but has only a memory requirement of $\mathcal{O}(1)$. Since LRP is a backpropagation-based method, gradient checkpointing (Chen et al., 2016) techniques can be applied. In checkpointing, LRP requires two forward and one backward pass, while the memory requirement scales logarithmic with the number of layers. In Appendix B.6, we benchmark energy, time and memory consumption of LRP against perturbation-based methods across context- and model-sizes.

Table 2. Computational and memory complexity of LRP-based and linear-time perturbation methods measured w.r.t. a single forward pass. N_L : number of layers, N_T : number of tokens

Methods	Computational Complexity	Memory Consumption
LRP Checkpointing	$\mathcal{O}(1)$	$\mathcal{O}(\sqrt{N_L})$
Perturbation (linear)	$\mathcal{O}(N_T)$	$\mathcal{O}(1)$

4.3. Understanding & Manipulating Neurons (Q3)

In our investigation, we use the Phi-1.5 model (Li et al., 2023), which has a transformer-based architecture with a next-word prediction objective. We obtain reference samples for each knowledge neuron by collecting the most activating sentences over the Wikipedia summary dataset (Scheepers, 2017).

To illustrate, we consider the prompt: ‘The ice bear lives in the’ which gives the corresponding prediction: ‘Arctic’. Using AttnLRP, we determine the most relevant layers for predicting ‘Arctic’ as well as the specific neurons within the FFN layers contributing to this prediction. Our analysis reveals that the most relevant neurons after the first three layers are predominantly situated within the middle layers. Notably, one standout neuron #3948 in layer 17 activates on reference samples about cold temperatures, as depicted in Figure 3. This observation is further validated by projecting the weight matrix of the second FFN layer onto the vocabulary. The neuron shifts the output distribution of the model to cold places, winter sports and animals living in cold regions.

Analogously, for the prompt ‘Children love to eat sugar and’ with the prediction ‘sweets’, the most relevant neuron’s (layer 18, neuron #5687) projection onto the vocabulary signifies a shift in the model’s focus towards the concept of candy, temptation and sweetness in the vocabulary space. We interact with the model by deactivating neuron #3848, and strongly amplifying the activation of neuron #5687 in the forward pass. This manipulation yields the following prediction change:

Prompt: Ice bears live in the
Prediction: sweet, sugary treats of the
candy store.

We further notice that neuron #4104 in layer 17 encodes for dryness, thirst and sand. Increasing its activation changes the output to ‘desert’ (illustrated in Figure 2).

With AttnLRP, we are able to trace the most important neurons in models with billions of parameters. This allows us to systematically navigate the latent space to enable targeted modifications to reduce the impact of certain concepts (for example, ‘coldness’) and enhance the presence of other concepts (for example, ‘dryness’), resulting in discernible output changes. Such an approach holds significant implications for transformer-based models, which have been difficult to manipulate and explain due to inherent opacity and size.

5. Conclusion

We have extended the Layer-wise Relevance Propagation framework to non-linear attention, proposing novel rules for the softmax and matrix-multiplication step and providing interpretations in terms of Deep Taylor Decomposition. Our AttnLRP method stands out due to its unique combination of simplicity, faithfulness, and efficiency. We demonstrate its applicability both for LLMs as well as ViTs, utilizing the stabilizing effect of the γ -rule. In contrast to other backpropagation-based approaches, AttnLRP enables the accurate attribution of neurons in latent space (also within the attention module), thereby introducing novel possibilities for real-time model interaction and interpretation.

Limitations & Open Problems

To enhance efficiency and reduce memory consumption, custom GPU kernels analogously to FlashAttention (Dao et al., 2022) should be implemented for LRP rules. Tuning the γ -parameter in ViTs to obtain faithful attributions is necessary.

References

- Abnar, S. and Zuidema, W. H. (2020). Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197.
- Achitbat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., and Lapuschkin, S. (2023). From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019.
- Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.-R., and Wolf, L. (2022). Xai for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning*, pages 435–451. PMLR.
- Arras, L., Osman, A., and Samek, W. (2022). Cleverxai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40.
- Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140.
- Baldazzi, D., Frean, M., Leary, L., Lewis, J., Ma, K. W.-D., and McWilliams, B. (2017). The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning*, pages 342–350. PMLR.
- Bills, S., Cammarata, N., Mossing, D., Tillman, H., et al. (2023). Language models can explain neurons in language models. URL <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html> (accessed: 14.05.2023).
- Blücher, S., Vielhaben, J., and Strothoff, N. (2024). Decoupling pixel flipping and occlusion strategy for consistent xai benchmarks. *arXiv preprint arXiv:2401.06654*.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660.
- Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. (2018). Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024*.

-
- Chefer, H., Gur, S., and Wolf, L. (2021a). Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406.
- Chefer, H., Gur, S., and Wolf, L. (2021b). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791.
- Chen, T., Xu, B., Zhang, C., and Guestrin, C. (2016). Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does bert look at? an analysis of bert’s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.
- Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. (2022). Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.
- Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. (2022). Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.
- Deb, M., Deisereth, B., Weinbach, S., Schramowski, P., and Kersting, K. (2023). Atman: Understanding transformer predictions through memory efficient attention manipulation. *arXiv preprint arXiv:2301.08110*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Ding, Y., Liu, Y., Luan, H., and Sun, M. (2017). Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159.
- Dombrowski, A.-K., Anders, C. J., Müller, K.-R., and Kessel, P. (2022). Towards robust explanations for deep neural networks. *Pattern Recognition*, 121:108194.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR*.
- Fatima, S. S., Wooldridge, M., and Jennings, N. R. (2008). A linear approximation method for the shapley value. *Artificial Intelligence*, 172(14):1673–1699.
- Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457.
- Fryer, D., Strümke, I., and Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*, 9:144352–144360.
- Geva, M., Caciularu, A., Wang, K., and Goldberg, Y. (2022). Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45.
- Geva, M., Schuster, R., Berant, J., and Levy, O. (2021). Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.
- Gildenblat, J. (2020. Accessed on Dec 01, 2023). Exploring explainability for vision transformers. <https://jacobjgil.github.io/deeplearning/vision-transformer-explainability>.
- Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., and Giannotti, F. (2018). Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunnasekar, S., and Lee, Y. T. (2023). Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.
- Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774.

-
- Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.
- Mao, C., Jiang, L., Dehghani, M., Vondrick, C., Sukthankar, R., and Essa, I. (2021). Discrete representations strengthen vision transformer robustness. In *International Conference on Learning Representations*.
- Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.
- Miglani, V., Yang, A., Markosyan, A., Garcia-Olano, D., and Kokhlikyan, N. (2023). Using captum to explain generative language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 165–173.
- Montavon, G., Binder, A., Lapuschkin, S., Samek, W., and Müller, K.-R. (2019). Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209.
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222.
- Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, 29.
- Pahde, F., Yolcu, G. Ü., Binder, A., Samek, W., and Lapuschkin, S. (2023). Optimizing explanations by network canonization and hyperparameter search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3818–3827.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.
- Radford, A., Jozefowicz, R., and Sutskever, I. (2017). Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.
- Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don’t know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (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*, pages 1135–1144. ACM.
- Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673.
- Scheepers, T. (2017). Improving the compositionality of word embeddings. Master’s thesis, Universiteit van Amsterdam, Science Park 904, Amsterdam, Netherlands.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.
- Shaham, U., Ivgi, M., Efrat, A., Berant, J., and Levy, O. (2023). Zeroscrolls: A zero-shot benchmark for long text understanding. *arXiv preprint arXiv:2305.14196*.
- Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR.
- Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR.
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.
- Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Voita, E., Ferrando, J., and Nalmpantis, C. (2023). Neurons in large language models: Dead, n-gram, positional. *arXiv preprint arXiv:2309.04827*.

Voita, E., Sennrich, R., and Titov, I. (2021). Analyzing the source and target contributions to predictions in neural machine translation. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, pages 1126–1140.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20.

Wikimedia Foundation (2023). Accessed on Dec 01, 2023). Wikimedia downloads. <https://dumps.wikimedia.org>.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference Computer Vision - ECCV 2014*, pages 818–833.

Zhang, B. and Sennrich, R. (2019). Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

Appendix

A. Appendix I: Methodological Details

This appendix provides further details on the methods presented in the paper. In particular, we focus on the AttnLRP method and provide implementation details and discuss the stability of the bias term and the noise problem in Vision Transformers. Finally, we provide proofs for the four propositions presented in the main paper.

A.1. Details on Baseline Methods

In the following, we present an overview of the baseline methods and their hyperparameter choices.

A.1.1. INPUT × GRADIENT

Gradients are one of the most straightforward approaches to depict how sensitive the trained model is with respect to each individual given feature (traditionally of the input space). By weighting the gradient with the input features, the model is locally linearized (Simonyan et al., 2014):

$$I \times G(\mathbf{x}) = \frac{\partial f_c(\mathbf{x})}{\partial \mathbf{x}} \times \mathbf{x} \quad (20)$$

Due to the gradient shattering effect (Balduzzi et al., 2017) which is a known phenomenon (especially in the ReLU-based CNNs), heatmaps generated by $I \times G$ are very noisy, making them in many cases not meaningful.

A.1.2. INTEGRATED GRADIENTS

To tackle the noisiness of $I \times G$, the idea to integrate gradients along a trajectory has been proposed. Here, the gradients of different (m) interpolated versions of the input \mathbf{x} , noted by \mathbf{x}' , are integrated as (Sundararajan et al., 2017):

$$\begin{aligned} IG(\mathbf{x}) &= (\mathbf{x} - \mathbf{x}') \int_{\alpha=0}^1 \frac{\partial f_j(\mathbf{x}' + \alpha \times (\mathbf{x} - \mathbf{x}'))}{\partial \mathbf{x}} d\alpha \\ &\approx (\mathbf{x} - \mathbf{x}') \sum_{k=1}^m \frac{\partial f_j(\mathbf{x}' + \frac{k}{m} \times (\mathbf{x} - \mathbf{x}'))}{\partial \mathbf{x}} \times \frac{1}{m} \end{aligned} \quad (21)$$

A.1.3. SMOOTHGRAD

A different technique towards the reduction of noisy gradients is smoothing the gradients (Smilkov et al., 2017) through generating (m) various samples in the neighborhood of input \mathbf{x} as $\mathbf{x}_\varepsilon = \mathbf{x} + \mathcal{N}(\mu, \sigma^2)$ and computing the average of all gradients:

$$\text{SmoothGrad}(\mathbf{x}) = \frac{1}{m} \sum_1^m \frac{\partial f_j(\mathbf{x}_\varepsilon)}{\partial \mathbf{x}_\varepsilon} \quad (22)$$

In this work, we set $\mu = 0$ and perform a hyperparameter search for σ to find the optimal parameter.

A.1.4. ATTENTION ROLLOUT

Self-Attention rollout (Abnar and Zuidema, 2020) capitalizes on the intrinsic nature of the attention weights matrix as a representative measure of token importance. It generates a $q \times k$ matrix where each row is normalized to form a probability distribution, representing the importance of each query token to all key tokens. To compute the relevance of hidden layer tokens (h) to the original input tokens (i), an iterative multiplication of the attention matrices on the left side is sufficient. To account for the residual connection through which the information of the previous tokens flows, an identity matrix \mathbf{I} is added:

$$\mathbf{R}_k^{h,i} = (\mathbf{I} + \mathbf{A}^{h,h}) \cdot \mathbf{R}_{k-1}^{h,i} \quad (23)$$

where $k = 0$ corresponds to the input layer and $R_{-1}^{h,i}$ is initialized with the identity matrix \mathbf{I} , h denotes the hidden feature space, and i stands for input dimension.

(Chefer et al., 2021a) build upon self-attention rollout and weights the attention matrix with the gradient. Additionally, the weighted attention map is denoised by computing the mean value of only positive values.

$$\bar{\mathbf{A}} = \mathbb{E}_b((\nabla \mathbf{A} \odot \mathbf{A})^+)$$

where \mathbb{E}_b denotes the expectation along the head dimension of the attention map. $\bar{\mathbf{A}}$ replaces $\mathbf{A}^{h,h}$ in (23). For encoder-decoder models, (Chefer et al., 2021a) present several additional considerations that are not mentioned here for brevity.

(Gildenblat, 2023) notes, that the rollout attributions can further be improved by discarding outlier values. For that, we define a discard threshold $dt \in [0, 1]$ used to compute the quantile $Q(dt)$, where dt represents the proportion of the data below the quantile e.g. with cumulative distribution function $P(\bar{\mathbf{A}} \leq Q(dt)) = dt$.

$$\bar{\mathbf{A}}_{m,n} = \begin{cases} 0 & \text{if } \bar{\mathbf{A}}_{m,n} > Q(dt) \\ \bar{\mathbf{A}}_{m,n} & \text{otherwise} \end{cases}$$

A.1.5. ATMAN

AtMan (Deb et al., 2023) perturbs the pre-softmax activations along the k -dimension:

$$\begin{aligned} \mathbf{H} &= \mathbf{Q} \cdot \mathbf{K}^T \\ \tilde{\mathbf{H}} &= \mathbf{H} \odot (\mathbf{1} - \mathbf{p}^i) \end{aligned}$$

where $\mathbf{H} \in \mathbb{R}^{b \times q \times k}$, and $\mathbf{1} \in [1]^{b \times q \times k}$ a matrix containing only 1. \mathbf{p}^i denotes a matrix $\in \mathbb{R}^{b \times q \times k}$ with

$$\mathbf{p}_{lmn}^i = \begin{cases} p & \text{for } n = i \\ 0 & \text{for } n \neq i \end{cases}$$

Thus, for a single token $i \in \{1, 2, \dots, N\}$, we suppress all values along the column/key-dimension with a suppression factor p . The suppression factor is a hyperparameter that must be tuned to the dataset and model. For ViTs, additional cosine similarities are computed to suppress correlated tokens as detailed in (Deb et al., 2023). For that, an additional hyperparameter denoted as t for threshold must be optimized in ViTs only.

A.2. DETAILS ON ATNLRP

This section provides more details on AttnLRP and justifies the specific parameter choices made in our work (e.g., use of γ -LRP in Vision Transformers).

A.2.1. CONSERVATION & NUMERICAL STABILITY OF BIAS TERMS

The relevance of a specific neuron or function, denoted as R_j , is computed by summing the contributions of the input variables, represented by $\mathbf{J}_{ji} x_i$, and adding the contribution of the bias term, represented by \tilde{b}_j .

$$R_j = \left(\sum_i \mathbf{J}_{ji} x_i + \tilde{b}_j \right) \frac{R_j}{f_j(\mathbf{x})}$$

The relevance of the bias term itself, denoted as R_b , is calculated as:

$$R_b = \tilde{b}_j \frac{R_j}{f_j(\mathbf{x})}$$

If we want to compute the relevance of the input variables R_i while ensuring strict adherence to the conservation property (3), we must exclude the bias term, so that it does not absorb part of the relevance R_b . There are three options: We can either completely omit the bias term, distribute its relevance value uniformly across the input variables, or apply the identity rule.

Omitting the bias term: In this case, the relevance propagation equation is:

$$R_i = \sum_j R_{i \leftarrow j} = \sum_j \mathbf{J}_{ji} x_i \frac{R_j}{\sum_i \mathbf{J}_{ji} x_i + \varepsilon}$$

Here, we no longer divide by the original function $f_j(\mathbf{x})$. However, it is important to ensure that no sign flips occur, as $\sum_i \mathbf{J}_{ji} x_i$ might have a different sign than $f_j(\mathbf{x})$.

Distributing the bias term: Alternatively, we can distribute the relevance value of the bias term uniformly across the input variables. The relevance propagation equation in this case is:

$$R_i = \sum_j R_{i \leftarrow j} = \sum_j \left(\mathbf{J}_{ji} x_i + \frac{\tilde{b}_j}{N} \right) \frac{R_j}{f_j(\mathbf{x}) + \varepsilon}$$

Here, N represents the number of input variables and $R_b = 0$. However, this approach may lead to numerical instabilities if the function $f_j(\mathbf{x}) = 0$ with $x_i = 0$. Then, the relevance message $R_{i \leftarrow j}$ is:

$$R_{i \leftarrow j} = \frac{\tilde{b}_j}{N} \frac{R_j}{\varepsilon}$$

Since ε is very small, this term explodes and causes numerical instabilities. Thus, functions that fulfill $f(0) = 0$ should preferably linearized such that the bias term does not cause numerical instabilities.

Applying the identity rule: We can apply the identity rule as follows:

$$R_i^{l-1} = R_i^l$$

However, we may encounter numerical instabilities, but these effects will only become visible in the next sequential propagation rule, not at this layer yet. For example in the softmax function, we may encounter a situation where a $x_i^{l-1} = 0$ but $f_i^l(\mathbf{x}) > 0$. If a non-zero relevance value from layer l is assigned to $f_i^l(\mathbf{x})$, then its relevance value is propagated through the identity rule to x_i^{l-1} . Assuming we apply the ε -rule after the identity rule, the relevance is given by:

$$R_{k \leftarrow i}^{(l-2, l-1)} = \mathbf{J}_{ik} x_k^{l-2} \frac{R_i^{l-1}}{0 + \varepsilon}$$

Note, that we divide by $x_i^{l-1} = f_i^{l-1} = 0$, which results in numerical instabilities. Further note, that these instabilities would not occur if $x_k^{l-2} = 0$ or $R_i^{l-1} = 0$, which is the case for a function fulfilling $f_i^l(0) = 0$.

In summary, omitting the bias term completely, distributing its relevance value uniformly across the input variables or applying the identity rule are possible approaches, but they have their considerations and potential challenges. Regarding the softmax non-linearity, (Voita et al., 2021) distributes the bias term equally on all input variables, while (Ding et al., 2017; Chefer et al., 2021b) apply the element-wise identity rule (9). Both variants can lead in this case to severe numerical instabilities.

A.2.2. HIGHLIGHTING THE DIFFERENCE BETWEEN VARIOUS LRP METHODS

In Table 3, we illustrate the different strategies employed for LRP in the past.

As discussed in Appendix A.2.1, distributing the bias term uniformly or applying the identity rule on the softmax function can lead to numerical instabilities that we also observed in our experiments.

Applying the ε -rule on bi-linear matrix multiplication does violate the conservation property (3) as proved in Lemma 3 of Chefer et al. (2021b). As a heuristic, (Chefer et al.,

2021b) applies a normalization step by dividing both arguments by the summation of its absolute values.

(Ali et al., 2022) regards the softmax output as constant and does not propagate relevance through it. This way, the matrix multiplication is not bi-linear anymore, but becomes linear, resulting in the attribution of only the value path. Consequently, the query and key matrices can no longer be attributed, which reduces the faithfulness and makes latent explanations in query and key matrices infeasible.

In Figure 5, we illustrate different attribution maps for all four options to handle the softmax function. The given section is from the Wikipedia article on Mount Everest. The model is expected to provide the next answer token for the question ‘How high did they climb in 1922? According to the text, the 1922 expedition reached 8,’. For the correctly predicted token 3 is the attribution computed.

While the relevance values for AttnLRP or CP-LRP are between $[-4, 4]$, distributing the bias uniformly on the input variables or applying the identity rule leads to an explosion of the relevances between $[-10^{15}, 10^{15}]$. As a consequence, the heatmaps resemble random noise. AttnLRP highlights the correct token the strongest, while CP-LRP focuses strongly on the start-of-sequence $\langle s \rangle$ token and exhibits more background noise e.g. irrelevant tokens such as ‘Context’, ‘attracts’, ‘Everest’ are highlighted, while AttnLRP does not highlight them or assigns negative relevance. In Appendix B.5, we compare also other baseline methods. Note, that the model attends to numerous tokens within the text which enables it to derive conclusions. Consequently, an attribution that reflects the model behavior will highlight more than just the single accurate answer token. The faithfulness experiments in Table 1 demonstrate, the AttnLRP captures the model reasoning most accurately.

A.2.3. TACKLING NOISE IN VISION TRANSFORMERS

Since backpropagation-based attributions utilize the gradient, they may produce noisy attributions in models with many layers, where gradient shattering and noisy gradients appear (Balduzzi et al., 2017; Dombrowski et al., 2022). Hence, various adaptions of the ε -LRP rule were developed to strengthen the signal-to-noise ratio by dampening counter-acting activations (Bach et al., 2015; Montavon et al., 2019). Here, we use the generalized γ -rule that encompasses all other proposed rules in the literature (Montavon et al., 2019). Let z_{ij} be the contribution of input i to output j , e.g. $\mathbf{W}_{ji}x_i$, and z_j the neuron output activation. Then depending on the sign of z_j :

$$R_{i \leftarrow j}^{(l, l+1)} = \begin{cases} \frac{z_{ij} + \gamma z_{ij}^+}{z_j + \gamma \sum_k z_{kj}^+} R_j^{l+1} & \text{if } z_j > 0 \\ \frac{z_{ij} + \gamma z_{ij}^-}{z_j + \gamma \sum_k z_{kj}^-} R_j^{l+1} & \text{else} \end{cases} \quad (24)$$

Table 3. Conceptual differences between various-LRP methods and their implications.

Methods	Softmax	Matrix Multiplication	Layer Normalization
(Ding et al., 2017)	Identity rule ⇒ unstable (Appendix A.2.1)	ε -LRP ⇒ violates conservation	not available
(Voita et al., 2021)	Taylor Decomposition at \mathbf{x} (distributes the bias uniformly) ⇒ unstable (Appendix A.2.1)	ε -LRP ⇒ violates conservation	Taylor Decomposition at \mathbf{x} (distributes the bias uniformly) ⇒ unstable (Appendix A.2.1)
(Chefer et al., 2021b)	Identity rule ⇒ unstable (Appendix A.2.1)	ε -LRP (with absolute normalization)	Identity rule
(Ali et al., 2022)	Regarded as constant ⇒ No attribution inside attention module	ε -LRP (ensures conservation)	Identity rule
AttnLRP	Taylor Decomposition at \mathbf{x} (with bias)	ε -LRP & uniform rule (ensures conservation)	Identity rule

with $\gamma \in \mathbb{R}^{>0}$, $(\cdot)^+ = \max(\cdot, 0)$ and $(\cdot)^- = \min(\cdot, 0)$. If $\gamma = \infty$, it will then be equivalent to the LRP- z^+ -rule, which is given as

$$R_{i \leftarrow j}^{(l, l+1)} = \frac{(w_{ij}x_i)^+}{z_j^+} R_j^{l+1} \quad (25)$$

by only taking into account positive contributions $z_j^+ = \sum_i (w_{ij}x_i)^+$ with $(\cdot)^+ = \max(0, \cdot)$. This is the case of softmax layers (of the attention layer) where only positive values are dealt with.

Remarkably, our observations reveal that attributions in LLMs demonstrate high sparsity and lack visible noise, while ViTs are susceptible to gradient shattering. We hypothesize that the discrete nature of the text domain may affect robustness (Mao et al., 2021). Therefore, we only apply the γ -rule in ViTs in the convolutional and linear FFN layers outside the attention module. To further increase the faithfulness, the γ -rule can be also applied on softmax layers. Since the output of the softmax is always greater than zero, we apply the simplified z^+ -rule (special case of γ -rule).

A.3. Proofs

In the following, we provide proofs for the rules presented in the main paper.

A.3.1. PROPOSITION 3.1: DECOMPOSING SOFTMAX

In this subsection, we demonstrate the decomposition of the softmax function by linearizing (4) it at \mathbf{x} . We begin by considering the softmax function:

$$s_j(\mathbf{x}) = \frac{e^{x_j}}{\sum_i e^{x_i}}$$

The derivative of the softmax has two cases, which depend on the output and input indices i and j :

$$\frac{\partial s_j}{\partial x_i} = \begin{cases} s_j(1 - s_j) & \text{for } i = j \\ -s_j s_i & \text{for } i \neq j \end{cases}$$

Consequently, a Taylor decomposition (4) yields:

$$f_j(\mathbf{x}) = s_j \left(x_j - \sum_i s_i x_i \right) + \tilde{b}_j$$

We differentiate between two cases, namely (i) when we attribute relevance from output j to input $i \neq j$ and (ii) when we attribute from output j to input $i = j$.

$$R_{i \leftarrow j} = \begin{cases} (x_i - s_i x_i) R_i^{l+1} & \text{for } i = j \\ -s_i x_i R_j^{l+1} & \text{for } i \neq j \end{cases}$$

Applying equation (5), we obtain:

$$R_i^l = \sum_j R_{i \leftarrow j} = x_i (R_i^{l+1} - s_i \sum_j R_j^{l+1})$$

A.3.2. PROPOSITION 3.2: DECOMPOSING MULTIPLICATION

The aim in this subsection is to decompose the multiplication of N input variables.

$$f_j(\mathbf{x}) = \prod_i^N x_i$$

We start by performing a Taylor Decomposition (4), then we derive the same decomposition with Shapley.

Taylor Decomposition: The derivative is

$$\frac{\partial f_j}{\partial x_i} = \prod_{k \neq i}^N x_k$$

Consequently, a Taylor decomposition (4) at \mathbf{x} yields

$$f_j(\mathbf{x}) = \sum_i^N \frac{\partial f_j}{\partial x_i} x_i + \tilde{b}_j = N \prod_k^N x_k + \tilde{b}_j = N f_j(\mathbf{x}) + \tilde{b}_j$$

We can either omit the bias term or equally distribute it on the input variables to strictly enforce the conservation property (3). Here, we demonstrate how to distribute the bias term uniformly.

$$R_{i \leftarrow j} = \left(f_j + \frac{\tilde{b}_j}{N} \right) \frac{R_j}{N f_j(\mathbf{x}) + \tilde{b}_j} = \frac{1}{N} R_j$$

For omitting the bias term, repeat the proof with $\tilde{b}_j = 0$.

Shapley: The Shapley value (Lundberg and Lee, 2017) is defined as:

$$\phi_i(f) = \sum_{\substack{S \subseteq N \\ i \notin S}} \frac{|S|!(N - |S| - 1)!}{N!} (f(S \cup \{i\}) - f(S)) \quad (26)$$

where $\phi_i(v)$ is the Shapley value of the feature i and value function f . N denotes the set of all features, and S denotes a feature subset (coalition).

With respect to multiplication, zero is the absorbing element. Hence, we choose zero as our baseline value, and the Shapley value function becomes:

$$\begin{aligned} f(S \cup \{i\}) &= \prod_k x_k \\ f(S) &= 0 \\ f(S \cup \{i\}) - f(S) &= \prod_k x_k \end{aligned}$$

The symmetry theorem (Fryer et al., 2021) of Shapley states that the contributions of two feature values i and l should be the same if they contribute equally to all possible coalitions

$$\begin{aligned} f(S \cup \{i\}) &= f(S \cup \{l\}) \\ \forall S \subseteq \{1, 2, \dots, N\} \setminus \{i, l\} \end{aligned}$$

then $\phi_i(f) = \phi_l(f)$. In addition, the efficiency theorem (Fryer et al., 2021) states that the output contribution is distributed equally amongst all features. Hence, the output contribution is equal to the sum of coalition values of all features i ,

$$\sum_i \phi_i(f) = f(N)$$

Both theorems are applicable and hence it follows:

$$\phi_i(f) = \frac{1}{N} f(N)$$

In the case of LRP, we identify $f(N)$ as R_j .

A.3.3. PROPOSITION 3.3: DECOMPOSING BI-LINEAR MATRIX MULTIPLICATION

Consider the equation for matrix multiplication, where we treat the terms as single input variables:

$$\mathbf{O}_{jp} = \sum_i (\mathbf{A}_{ji} \mathbf{V}_{ip})$$

In this case, the function already is in the form of an additive decomposition (1). Therefore,

$$R_{jip}(\mathbf{A}_{ji}, \mathbf{V}_{ip}) = \mathbf{A}_{ji} \mathbf{V}_{ip} \frac{R_{jp}}{\mathbf{O}_{jp}}$$

Next, we decompose the individual terms using the derived rule from the previous Section A.3.2

$$R_{ji} = \sum_p R_{ji \leftarrow jip} = \sum_p \frac{1}{2} R_{jip} = \sum_p \mathbf{A}_{ji} \mathbf{V}_{ip} \frac{R_{jp}}{2 \mathbf{O}_{jp}}$$

The proof for \mathbf{V}_{ip} follows a similar approach.

A.3.4. PROPOSITION 3.4: LAYER NORMALIZATION

Consider layer normalization of the form

$$f_j(\mathbf{x}) = \frac{x_j}{g(\mathbf{x})}$$

where $g(\mathbf{x}) = \sqrt{\text{Var}[\mathbf{x}] + \varepsilon}$ or $g(\mathbf{x}) = \sqrt{\frac{1}{N} \sum_k x_k^2 + \varepsilon}$.

The derivative is

$$\frac{\partial f_j}{\partial x_i} = \frac{1}{g(\mathbf{x})^2} \begin{cases} g(\mathbf{x}) - x_j \frac{\partial g(\mathbf{x})}{\partial x_i} & \text{for } i = j \\ -x_j \frac{\partial g(\mathbf{x})}{\partial x_i} & \text{for } i \neq j \end{cases} \quad (27)$$

In LayerNorm (Ba et al., 2016), we assume for simplicity $\mathbb{E}[\mathbf{x}] = 0$, then the partial derivative simplifies to

$$\begin{aligned} \mathbb{V}[\mathbf{x}] &= \mathbb{E}[\mathbf{x}^2] - \mathbb{E}[\mathbf{x}]^2 = \mathbb{E}[\mathbf{x}^2] \\ \frac{\partial \mathbb{V}[\mathbf{x}]}{\partial x_i} &= \frac{2}{N} x_i \end{aligned}$$

Further, the partial derivative of RMSNorm (Zhang and Sennrich, 2019) is

$$\frac{\partial \text{RMSNorm}}{\partial x_i} = \frac{x_i}{\sqrt{N \sum_k x_k}}$$

At reference point $\tilde{x}_i = 0$, the diagonal elements in equation (27) $i \neq j$ are zero, yielding the Taylor decomposition:

$$f_j(\mathbf{x}) = \frac{\partial f_j}{\partial x_i} \Big|_{\tilde{x}_i=0} x_i + \tilde{b}_j = \frac{x_i}{\varepsilon} + \tilde{b}_j$$

To enforce a strict notion of the conservation property (3), the bias term \tilde{b}_j can be excluded or evenly distributed across the input variables. Because we have only a single input variable, the bias can be considered as part of x_i .

$$R_i = \left(\frac{x_i}{\varepsilon} + \tilde{b}_j \right) \frac{R_j}{\frac{x_i}{\varepsilon} + \tilde{b}_j} \quad (28)$$

Since there is only one input variable and one output, the decomposition is equivalent to the identity function, as discussed in the Section 3.2.2 about component-wise nonlinearities. Thus, we conclude that the identity rule applies in this case.

$$R_i^{l-1} = R_i^l \quad (29)$$

Note, that this rule is numerically stable because $f_j(0) = 0$ as discussed in Section A.2.1.

B. Appendix II: Experimental Details

In the following sections, we provide additional details about the experiments performed.

B.1. Models and Datasets

For ImageNet faithfulness, we utilized the pretrained Vision Transformer 16-B weights of the PyTorch model zoo (Paszke et al., 2019). We randomly selected 3200 samples such that the standard error of mean converges to below 1% of the mean value.

For Wikipedia, SQuAD v2 and IMDB faithfulness, we evaluated the pretrained Llama 2-7b and Flan-T5 weights hosted on huggingface (Wolf et al., 2019) on 4000 randomly selected validation dataset samples (fixed set for all baselines). Both models are casted to the Brain Floating Point (BFLOAT16) half-precision format to save memory consumption and to simulate real-world scenarios. Further, for Wikipedia next word prediction we evaluated the model on a context size of 512 (from beginning of article until context length is reached), while the context size in SQuAD v2 varies between 169 to 4060. Although Flan-T5 was trained on a smaller context size of 2000 tokens, the relative positional encoding allows it to handle longer context sizes with at least 8192 tokens (Shaham et al., 2023). Additionally for SQuAD v2 faithfulness and IoU, we utilized the following prompt:

```
Context: [text of dataset sample]
Question: [question of dataset sample]
Answer:
```

In order to prevent out-of-distribution samples, we only perturb the tokens inside the context [text of dataset sample]. Since the length of the SQuAD v2 questions varies and the length of the validation dataset is small, the standard error of mean is higher in Table 1.

For IMDB, we added a last linear layer to Llama 2-7b and finetuned the model, which achieves 93% accuracy on the validation dataset.

If we encountered NaN values for a sample, we removed it from the evaluation. This happened for Grad \times AttnRollout and AtMan in the Wikipedia dataset. However, the standard error of the mean remains small, as can be seen in Table 1.

B.2. Input Perturbation Metrics

In the following, we summarize the perturbation process introduced by (Samek et al., 2017) in a condensed manner.

Given an attributions map $R_i^l(x_i)$ per input features $\mathbf{x} = \{x_i\}_{i=1}^N$ in layer l . \mathcal{H} denotes a set of relevance values for all input features x_i :

$$\mathcal{H} = (R_0^0(x_0), R_1^0(x_1), \dots, R_{N-1}^0(x_{N-1})) \quad (30)$$

Then, the *flipping* perturbation process iteratively substitutes input features with a baseline value $\mathbf{b} \in \mathbb{R}^N$ (the baseline might be zero, noise generated from a Gaussian distribution, or pixels of a black image in the vision task). Another reverse equivalent variant, referred as *insertion*, begins with a baseline \mathbf{b} and reconstructs the input \mathbf{x} stepwise. The function performing the perturbation is denoted by \mathbf{g}^F for flipping and \mathbf{g}^I for insertion. The perturbation procedure is either conducted in a MoRF (Most Relevant First) or LeRF (Least Relevant First) manner based on the sorted members of \mathcal{H} . Regardless of the replacement function, the MoRF and LeRF perturbation processes can be defined as recursive formulas at step $k = \{0, 1, \dots, N - 1\}$:

$$\text{MoRF Pert. Process} = \begin{cases} \mathbf{x}_{MoRF}^0 = \mathbf{x} \\ \mathbf{x}_{MoRF}^k = \mathbf{g}^{(F|I)}(\mathbf{x}_{MoRF}^{k-1}, \mathbf{b}) \\ \mathbf{x}_{MoRF}^{N-1} = \mathbf{b} \end{cases}$$

where \mathbf{x}_{MoRF}^k denotes the perturbed input feature \mathbf{x} at step k in MoRF process.

$$\text{LeRF Pert. Process} = \begin{cases} \mathbf{x}_{LeRF}^0 = \mathbf{b} \\ \mathbf{x}_{LeRF}^k = \mathbf{g}^{(F|I)}(\mathbf{x}_{LeRF}^{k-1}, \mathbf{b}) \\ \mathbf{x}_{LeRF}^{N-1} = \mathbf{x} \end{cases}$$

where \mathbf{x}_{LeRF}^k denotes the perturbed input feature \mathbf{x} at step k in LeRF process.

Results of these processes are perturbed input sets of $\mathcal{X}_{MoRF}^F = (\mathbf{x}_{MoRF}^0, \mathbf{x}_{MoRF}^1, \dots, \mathbf{x}_{MoRF}^{N-1})$ and $\mathcal{X}_{LeRF}^F = (\mathbf{x}_{LeRF}^0, \mathbf{x}_{LeRF}^1, \dots, \mathbf{x}_{LeRF}^{N-1})$. It is notable that by using \mathbf{g}^I , $\mathcal{X}_{LeRF}^I = \mathcal{X}_{MoRF}^F$ and vice versa. By feeding these sets to the model and computing the corresponding logit output f_j , a curve will be induced and consequently the area under the curve can be calculated:

$$A_{MoRF}^F = A_{LeRF}^I = \frac{1}{N} \sum_{k=0}^{N-1} f_j(\mathbf{x}_{MoRF}^k) \quad (31)$$

where $\mathbf{x}_{MoRF}^k \in \mathcal{X}_{MoRF}^F$ or $\mathbf{x}_{MoRF}^k \in \mathcal{X}_{LeRF}^I$.

By using \mathcal{X}_{MoRF}^I , $A_{MoRF}^I = A_{LeRF}^F$ can be computed similarly. A faithful explainer results in a low value of A_{MoRF}^F or A_{LeRF}^I . Further, a faithful explainer is expected to have large A_{LeRF}^F or A_{MoRF}^I values.

Ultimately to reduce introducing out-of-distribution manipulations and the sensitivity towards the chosen baseline value, the work of (Blücher et al., 2024) proposes to leverage both insights and to obtain a robust measure as

$$\begin{aligned} \Delta A^F &= A_{LeRF}^F - A_{MoRF}^F \\ \Delta A^I &= A_{MoRF}^I - A_{LeRF}^I \end{aligned}$$

where a higher score signifies a more faithful explainer.

We performed all faithfulness perturbations with a baseline value of zero. In the case of LLMs, we aggregated the relevance for each token and flipped the entire embedding vector of input tokens to the baseline value. For ViTs, we used the relevances of the input pixels and flipped input pixels to the baseline value.

B.3. Hyperparameter search for Baselines

As described in Section A.1, several baseline attribution methods have hyperparameters that must be tuned to the datasets. The hyperparameters of SmoothGrad ($\sigma \in [0.01, 0.25]$), AtMan (suppression value $\in [0.1, 1.0]$) and threshold $\in [0, 1.0]$), AttnRoll (discard threshold $\in [0.90, 1.00]$), and G×AttnRoll (discard threshold $\in [0.90, 1.00]$) need to be optimized. The best searched hyperparameters according to the perturbation experiments are available in the captions of Tables 5, 6, 7, 8.

B.4. LRP Composites for ViT

Applying the ε -rule on all linear layers inside LLMs is sufficient to obtain faithful and noise-free attributions. However, for the Vision Transformers, we apply the γ -rule on all linear layers (including the convolutional layers) outside the attention module. Since the γ -rule has a hyperparameter, the work (Pahde et al., 2023) proposed to tune the parameter using a grid-search. This optimization search (or in LRP known as composite search) is computational highly demanding.

The Vision Transformer consists of many linear layers. Our proposed approach is to use different γ values across different layer types. According to (Vaswani et al., 2017) the attention module consists of several linear layers which we refer to as *LinearInputProjection*:

$$\begin{aligned} \mathbf{Q} &= \mathbf{W}_q \mathbf{X} + \mathbf{b}_q \\ \mathbf{K} &= \mathbf{W}_k \mathbf{X} + \mathbf{b}_k \\ \mathbf{V} &= \mathbf{W}_v \mathbf{X} + \mathbf{b}_v \end{aligned}$$

In the attention layer, after the softmax (11), there exists another linear layer performing the output projection back into the residual stream, denoted as *LinearOutputProjection*:

$$\mathbf{y} = \mathbf{W}_o \mathbf{O} + \mathbf{b}_o$$

The other layers in the whole network, will be referred to as *Linear*.

The perturbation experiment had been conducted over these layers using different types of rules including Epsilon, ZPlus, Gamma, and AlphaBeta (with $\alpha = 2$ and $\beta = 1$ according to (Montavon et al., 2019)).

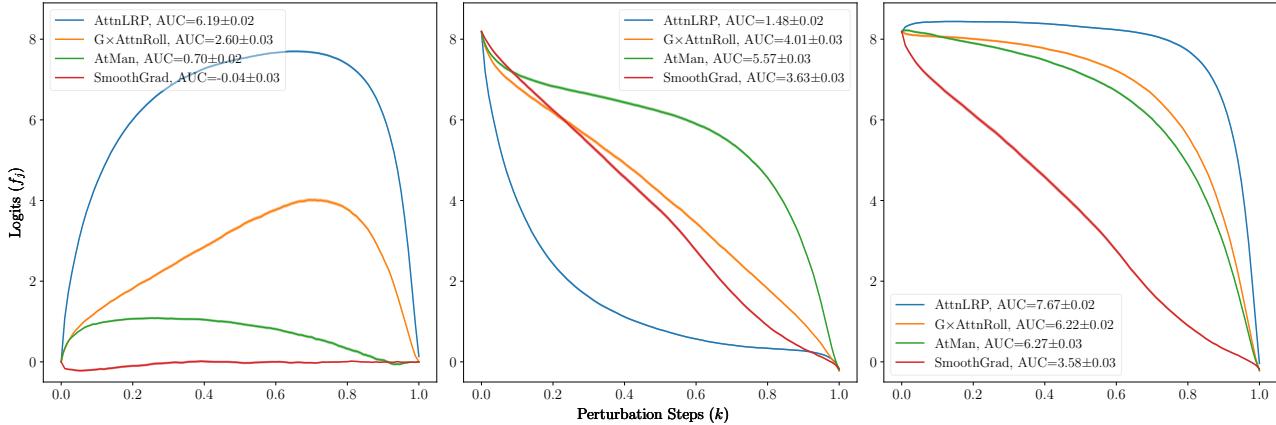


Figure 4. Comparison of the AttnLRP (ours) with the γ -rule, Grad \times AttnRoll (Chefer et al., 2021a), AtMan (Deb et al., 2023), and SmoothGrad (Smilkov et al., 2017) techniques through the perturbation experiment (faithfulness) on the ViT-B-16 using 3200 random samples of ImageNet. From left to right, the plots correspond to $f_j(\mathcal{X}_{LeRF}^F) - f_j(\mathcal{X}_{MoRF}^F)$ (large area is good), $f_j(\mathcal{X}_{MoRF}^F)$ (steep decline is good), and $f_j(\mathcal{X}_{LeRF}^F)$ (slow decline is good). “AUC” denotes the Area under Curve.

The most faithful composite, that we obtained for AttnLRP and CP-LRP, is in Table 4. More details over the statistics of the conducted experiments are available in Figures 12, 13, 14, 15, 16.

B.5. Attributions on SQuAD v2

In Figure 7, we illustrate attributions for different state-of-the-art methods computed for the first token of the answer (highlighted in red). The Flan-T5 model correctly predicts 182 million, where the first explained token is 18.

Gradient-based methods such as G \times I, SmoothGrad, IG or Grad-CAM are noisy and not really informative. While the heatmaps of I \times G or Grad-CAM seem to be inverted, we experimented with inverting the attributions on a subset, however we did not notice improvement and applied the rules with their original definition. Grad \times Attn Rollout suffers from background noise. AtMan produces highly sparse attributions, assigning large positive relevance to the answer token 18, however, also assigning a similar amount of relevance to the token much, which is part of the question. AttnLRP and CP-LRP identify the token 18 as being the most relevant token and also relate it (by assigning positive and negative relevance) to other information in the text such as 27.7, 132 or average. We conjecture that such targeted contrasting reflects the reasoning process of the model (*e.g.*, is necessary to distinguish between related questions about how many tons are blown out vs. how many tons remain on the ground). A systematic analysis of these effects remain an interesting topic for future work. The similarity between AttnLRP and CP-LRP in Flan-T5 are in line with the quantitative evaluation from Table 1, which showed a small, but consistent advantage of AttnLRP over CP-LRP wrt. faithfulness, while in Llama

2, AttnLRP substantially outperforms. For comparison, we also visualize a random attribution with Gaussian noise.

B.6. Benchmarking Cost, Time and Memory Consumption

We benchmark the runtime and peak GPU memory consumption for computing a single attribution for Llama 2 with batch size 1 on a node with four A100-SXM4 40GB, 512 GB CPU RAM and 32 AMD EPYC 73F3 3.5 GHz. Because AtMan, LRP and AttnRollout-variants need access to the attention weights, we did not use flash-attention (Dao et al., 2022).

To calculate energy cost, we assume a price of 0.16 \$ per kWh of energy, and that a single A100 GPU consumes on average 130W. Figure 8 depicts the cost, the runtime and peak GPU memory consumption. Since perturbation-based methods are memory efficient, a 70b model with full context size of 4096 is attributable. However, LRP with checkpointing requires more memory than a node can supply.

B.7. Attributions of Knowledge Neurons

Figure 9, 10 and 11 illustrate the top 10 sentences in the Wikipedia summary dataset that maximally activate a knowledge neuron. We applied AttnLRP to highlight the tokens inside these reference samples. We observe that knowledge neurons exhibit remarkable disentanglement, *e.g.*, neuron #256 of layer 18 shown in Figure 9 seems to encode concepts related to transport systems (railways in particular), while neuron #2207 of layer 20 shown in Figure 10 seems to encode the concept teacher, in particular a teacher, in an unusual context (*e.g.*, inappropriate behavior, sexual misconduct). The degree of disentangle-

Table 4. Proposed composite for the AttnLRP and CP-LRP methods used for the Vision Transformer.

Layer Type	Rule Proposed
Convolution	Gamma($\gamma = 0.25$)
Linear	Gamma($\gamma = 0.05$)
LinearInputProjection	Epsilon
LinearOutputProjection	Epsilon

AttnLRP

<s> Context: Mount Everest attracts many climbers, including highly experienced mountaineers. There are two main climbing routes, one approaching the summit from the southeast in Nepal (known as the standard route) and the other from the north in Tibet. While not posing substantial technical climbing challenges on the standard route, Everest presents dangers such as altitude sickness, weather, and wind, as well as hazards from avalanches and the Khumbu Icefall. As of November 2022, 310 people have died on Everest. Over 200 bodies remain on the mountain and have not been removed due to the dangerous conditions. The first recorded efforts to reach Everest's summit were made by British mountaineers. As Nepal did not allow foreigners to enter the country at the time, the British made several attempts on the north ridge route from the Tibetan side. After the first reconnaissance expedition by the British in 1921 reached 7,000 m (22,970 ft) on the North Col, the 1922 expedition pushed the north ridge route up to 8,320 m (27,300 ft), marking the first time a human had climbed above 8,000 m (26,247 ft). The 1924 expedition resulted in one of the greatest mysteries on Everest to this day: George Mallory and Andrew Irvine made a final summit attempt on 8 June but never returned, sparking debate as to whether they were the first to reach the top. Tenzing Norgay and Edmund Hillary made the first documented ascent of Everest in 1953, using the southeast ridge route. Norgay had reached 8,595 m (28,199 ft) the previous year as a member of the 1952 Swiss expedition. The Chinese mountaineering team of Wang Fuzhou, Gponpo, and Qu Yinhua made the first reported ascent of the peak from the north ridge on 25 May 1960. Question: How high did they climb in 1922? According to the text, the 1922 expedition reached 8.

CP-LRP

<s> Context: Mount Everest attracts many climbers, including highly experienced mountaineers. There are two main climbing routes, one approaching the summit from the southeast in Nepal (known as the standard route) and the other from the north in Tibet. While not posing substantial technical climbing challenges on the standard route, Everest presents dangers such as altitude sickness, weather, and wind, as well as hazards from avalanches and the Khumbu Icefall. As of November 2022, 310 people have died on Everest. Over 200 bodies remain on the mountain and have not been removed due to the dangerous conditions. The first recorded efforts to reach Everest's summit were made by British mountaineers. As Nepal did not allow foreigners to enter the country at the time, the British made several attempts on the north ridge route from the Tibetan side. After the first reconnaissance expedition by the British in 1921 reached 7,000 m (22,970 ft) on the North Col, the 1922 expedition pushed the north ridge route up to 8,320 m (27,300 ft), marking the first time a human had climbed above 8,000 m (26,247 ft). The 1924 expedition resulted in one of the greatest mysteries on Everest to this day: George Mallory and Andrew Irvine made a final summit attempt on 8 June but never returned, sparking debate as to whether they were the first to reach the top. Tenzing Norgay and Edmund Hillary made the first documented ascent of Everest in 1953, using the southeast ridge route. Norgay had reached 8,595 m (28,199 ft) the previous year as a member of the 1952 Swiss expedition. The Chinese mountaineering team of Wang Fuzhou, Gponpo, and Qu Yinhua made the first reported ascent of the peak from the north ridge on 25 May 1960. Question: How high did they climb in 1922? According to the text, the 1922 expedition reached 8,

Softmax Distribute Bias

<s> Context: Mount Everest attracts many climbers, including highly experienced mountaineers. There are two main climbing routes, one approaching the summit from the southeast in Nepal (known as the standard route) and the other from the north in Tibet. While not posing substantial technical climbing challenges on the standard route, Everest presents dangers such as altitude sickness, weather, and wind, as well as hazards from avalanches and the Khumbu Icefall. As of November 2022, 310 people have died on Everest. Over 200 bodies remain on the mountain and have not been removed due to the dangerous conditions. The first recorded efforts to reach Everest's summit were made by British mountaineers. As Nepal did not allow foreigners to enter the country at the time, the British made several attempts on the north ridge route from the Tibetan side. After the first reconnaissance expedition by the British in 1921 reached 7,000 m (22,970 ft) on the North Col, the 1922 expedition pushed the north ridge route up to 8,320 m (27,300 ft), marking the first time a human had climbed above 8,000 m (26,247 ft). The 1924 expedition resulted in one of the greatest mysteries on Everest to this day: George Mallory and Andrew Irvine made a final summit attempt on 8 June but never returned, sparking debate as to whether they were the first to reach the top. Tenzing Norgay and Edmund Hillary made the first documented ascent of Everest in 1953, using the southeast ridge route. Norgay had reached 8,595 m (28,199 ft) the previous year as a member of the 1952 Swiss expedition. The Chinese mountaineering team of Wang Fuzhou, Gponpo, and Qu Yinhua made the first reported ascent of the peak from the north ridge on 25 May 1960. Question: How high did they climb in 1922? According to the text, the 1922 expedition reached 8,

Softmax Identity Rule

<s> Context: Mount Everest attracts many climbers, including highly experienced mountaineers. There are two main climbing routes, one approaching the summit from the southeast in Nepal (known as the standard route) and the other from the north in Tibet. While not posing substantial technical climbing challenges on the standard route, Everest presents dangers such as altitude sickness, weather, and wind, as well as hazards from avalanches and the Khumbu Icefall. As of November 2022, 310 people have died on Everest. Over 200 bodies remain on the mountain and have not been removed due to the dangerous conditions. The first recorded efforts to reach Everest's summit were made by British mountaineers. As Nepal did not allow foreigners to enter the country at the time, the British made several attempts on the north ridge route from the Tibetan side. After the first reconnaissance expedition by the British in 1921 reached 7,000 m (22,970 ft) on the North Col, the 1922 expedition pushed the north ridge route up to 8,320 m (27,300 ft), marking the first time a human had climbed above 8,000 m (26,247 ft). The 1924 expedition resulted in one of the greatest mysteries on Everest to this day: George Mallory and Andrew Irvine made a final summit attempt on 8 June but never returned, sparking debate as to whether they were the first to reach the top. Tenzing Norgay and Edmund Hillary made the first documented ascent of Everest in 1953, using the southeast ridge route. Norgay had reached 8,595 m (28,199 ft) the previous year as a member of the 1952 Swiss expedition. The Chinese mountaineering team of Wang Fuzhou, Gponpo, and Qu Yinhua made the first reported ascent of the peak from the north ridge on 25 May 1960. Question: How high did they climb in 1922? According to the text, the 1922 expedition reached 8,

Figure 5. Comparison of four different LRP variants computed on a Llama 2-7b model. The given section is from the Wikipedia article on Mount Everest. The model is expected to provide the next answer token for the question ‘How high did they climb in 1922? According to the text, the 1922 expedition reached 8,’. For the correctly predicted token 3 is the attribution computed. Distributing the bias uniformly on the input variables (Softmax Distribute Bias) or applying the identity rule (Softmax Identity Rule) leads to numerical instabilities. AttnLRP highlights the correct token the strongest, while CP-LRP focuses strongly on the start-of-sequence <s> token and exhibits more background noise e.g. irrelevant tokens such as ‘Context’, ‘attracts’, ‘Everest’ are highlighted, while AttnLRP does not highlight them or assigns negative relevance.

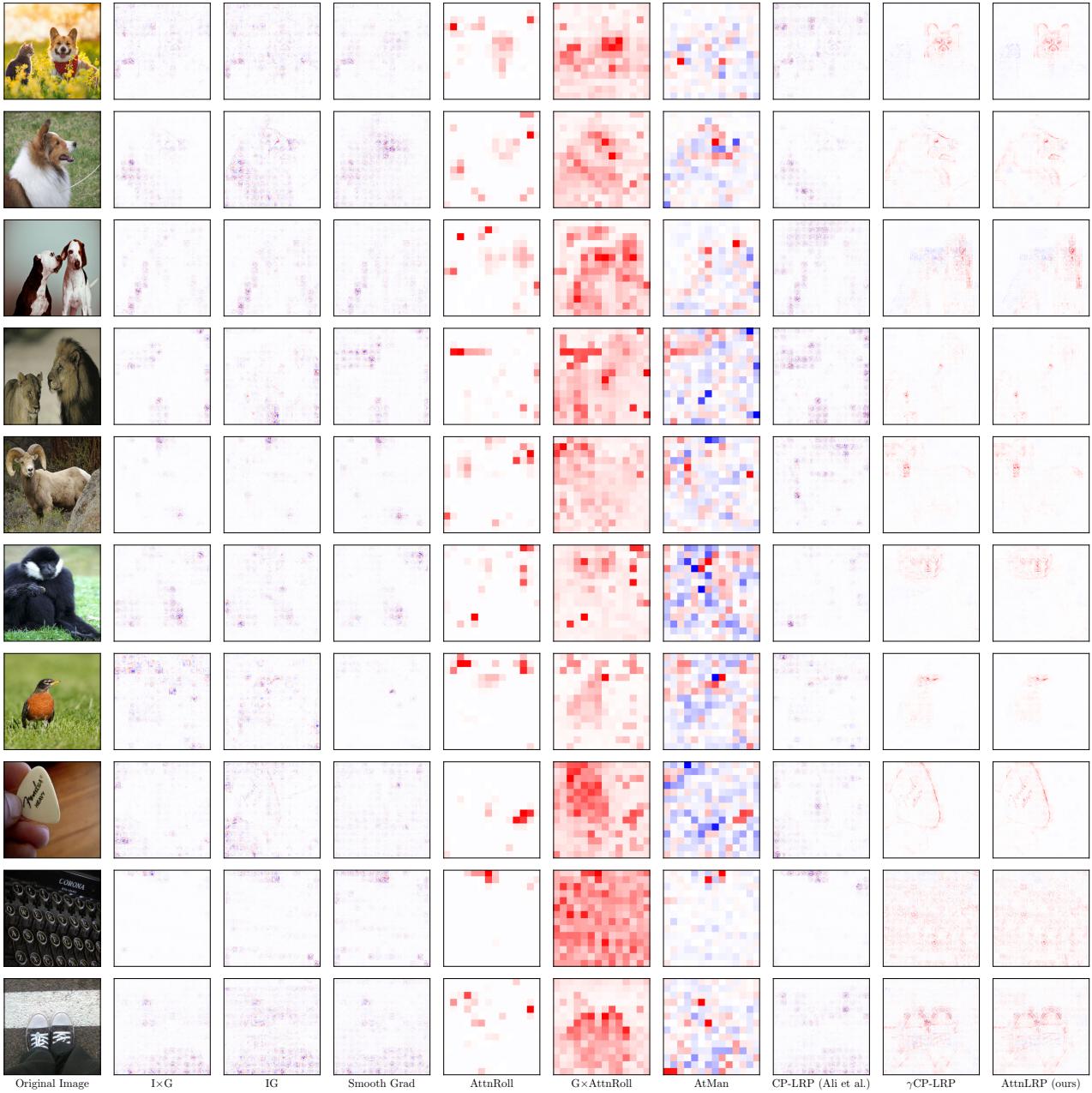


Figure 6. Explanation heatmaps of the methods used for the perturbation experiments on the Vision Transformer. A checkerboard effect is visible for almost every method, especially in AttnRoll (Abnar and Zuidema, 2020), G \times AttnRoll (Chefer et al., 2021a), and AtMan (Deb et al., 2023). We improve upon CP-LRP (Ali et al., 2022) by applying the γ -rule as described in B.4. While qualitatively CP-LRP (with γ extension for ViT) and AttnLRP give similar explanations, quantitative results in Table 1 show a consistent improvement of AttnLRP over CP-LRP in terms of faithfulness. Moreover, the detailed investigation of the processes within the attention model can be investigated with AttnLRP only, while it is not possible with CP-LRP. We leave these further explorations for future work. The quantitative experiments (faithfulness test) validate this observation, see Table 5.

Question: How many tons of dust are blown out of the Sahara each year?

Answer: 182 million

AttnLRP

Context: NASA's CALIPSO satellite has measured the amount of dust transported by wind from the Sahara to the Amazon: an average 182 million tons of dust are windblown out of the Sahara each year, at 15 degrees west longitude, across 1,600 miles (2,600 km) over the Atlantic Ocean (some dust falls into the Atlantic), then at 35 degrees West longitude at the eastern coast of South America, 27.7 million tons (15%) of dust fall over the Amazon basin, 132 million tons of dust remain in the air, 43 million tons of dust are windblown and falls on the Caribbean Sea, past 75 degrees west longitude. Question: How much dust is blown out of the Sahara each year? Answer:</s>

Gradient x Input

Context: NASA's CALIPSO satellite has measured the amount of dust transported by wind from the Sahara to the Amazon: an average 182 million tons of dust are windblown out of the Sahara each year, at 15 degrees west longitude, across 1,600 miles (2,600 km) over the Atlantic Ocean (some dust falls into the Atlantic), then at 35 degrees West longitude at the eastern coast of South America, 27.7 million tons (15%) of dust fall over the Amazon basin, 132 million tons of dust remain in the air, 43 million tons of dust are windblown and falls on the Caribbean Sea, past 75 degrees west longitude. Question: How much dust is blown out of the Sahara each year? Answer:</s>

AtMan

Context: NASA's CALIPSO satellite has measured the amount of dust transported by wind from the Sahara to the Amazon: an average 182 million tons of dust are windblown out of the Sahara each year, at 15 degrees west longitude, across 1,600 miles (2,600 km) over the Atlantic Ocean (some dust falls into the Atlantic), then at 35 degrees West longitude at the eastern coast of South America, 27.7 million tons (15%) of dust fall over the Amazon basin, 132 million tons of dust remain in the air, 43 million tons of dust are windblown and falls on the Caribbean Sea, past 75 degrees west longitude. Question: How much dust is blown out of the Sahara each year? Answer:</s>

Grad-CAM

Context: NASA's CALIPSO satellite has measured the amount of dust transported by wind from the Sahara to the Amazon: an average 182 million tons of dust are windblown out of the Sahara each year, at 15 degrees west longitude, across 1,600 miles (2,600 km) over the Atlantic Ocean (some dust falls into the Atlantic), then at 35 degrees West longitude at the eastern coast of South America, 27.7 million tons (15%) of dust fall over the Amazon basin, 132 million tons of dust remain in the air, 43 million tons of dust are windblown and falls on the Caribbean Sea, past 75 degrees west longitude. Question: How much dust is blown out of the Sahara each year? Answer:</s>

Integrated Gradient

Context: NASA's CALIPSO satellite has measured the amount of dust transported by wind from the Sahara to the Amazon: an average 182 million tons of dust are windblown out of the Sahara each year, at 15 degrees west longitude, across 1,600 miles (2,600 km) over the Atlantic Ocean (some dust falls into the Atlantic), then at 35 degrees West longitude at the eastern coast of South America, 27.7 million tons (15%) of dust fall over the Amazon basin, 132 million tons of dust remain in the air, 43 million tons of dust are windblown and falls on the Caribbean Sea, past 75 degrees west longitude. Question: How much dust is blown out of the Sahara each year? Answer:</s>

CP-LRP

Context: NASA's CALIPSO satellite has measured the amount of dust transported by wind from the Sahara to the Amazon: an average 182 million tons of dust are windblown out of the Sahara each year, at 15 degrees west longitude, across 1,600 miles (2,600 km) over the Atlantic Ocean (some dust falls into the Atlantic), then at 35 degrees West longitude at the eastern coast of South America, 27.7 million tons (15%) of dust fall over the Amazon basin, 132 million tons of dust remain in the air, 43 million tons of dust are windblown and falls on the Caribbean Sea, past 75 degrees west longitude. Question: How much dust is blown out of the Sahara each year? Answer:</s>

Grad-weighted Attention Rollout

Context: NASA's CALIPSO satellite has measured the amount of dust transported by wind from the Sahara to the Amazon: an average 182 million tons of dust are windblown out of the Sahara each year, at 15 degrees west longitude, across 1,600 miles (2,600 km) over the Atlantic Ocean (some dust falls into the Atlantic), then at 35 degrees West longitude at the eastern coast of South America, 27.7 million tons (15%) of dust fall over the Amazon basin, 132 million tons of dust remain in the air, 43 million tons of dust are windblown and falls on the Caribbean Sea, past 75 degrees west longitude. Question: How much dust is blown out of the Sahara each year? Answer:</s>

SmoothGrad

Context: NASA's CALIPSO satellite has measured the amount of dust transported by wind from the Sahara to the Amazon: an average 182 million tons of dust are windblown out of the Sahara each year, at 15 degrees west longitude, across 1,600 miles (2,600 km) over the Atlantic Ocean (some dust falls into the Atlantic), then at 35 degrees West longitude at the eastern coast of South America, 27.7 million tons (15%) of dust fall over the Amazon basin, 132 million tons of dust remain in the air, 43 million tons of dust are windblown and falls on the Caribbean Sea, past 75 degrees west longitude. Question: How much dust is blown out of the Sahara each year? Answer:</s>

Attention Rollout

Context: NASA's CALIPSO satellite has measured the amount of dust transported by wind from the Sahara to the Amazon: an average 182 million tons of dust are windblown out of the Sahara each year, at 15 degrees west longitude, across 1,600 miles (2,600 km) over the Atlantic Ocean (some dust falls into the Atlantic), then at 35 degrees West longitude at the eastern coast of South America, 27.7 million tons (15%) of dust fall over the Amazon basin, 132 million tons of dust remain in the air, 43 million tons of dust are windblown and falls on the Caribbean Sea, past 75 degrees west longitude. Question: How much dust is blown out of the Sahara each year? Answer:</s>

Random

Context: NASA's CALIPSO satellite has measured the amount of dust transported by wind from the Sahara to the Amazon: an average 182 million tons of dust are windblown out of the Sahara each year, at 15 degrees west longitude, across 1,600 miles (2,600 km) over the Atlantic Ocean (some dust falls into the Atlantic), then at 35 degrees West longitude at the eastern coast of South America, 27.7 million tons (15%) of dust fall over the Amazon basin, 132 million tons of dust remain in the air, 43 million tons of dust are windblown and falls on the Caribbean Sea, past 75 degrees west longitude. Question: How much dust is blown out of the Sahara each year? Answer:</s>

Figure 7. Evaluation on the SQuAD v2: We compute attributions for different state-of-the-art methods on the first token of the answer (highlighted in red). Gradient-based methods such as G×I, SmoothGrad, IG or Grad-CAM are noisy. Grad×Attn Rollout suffers from background noise. AtMan produces highly sparse attributions, assigning an equal amount of relevance to a token, which is part of the question, as to token 18. CP-LRP has a different weighting of the tokens e.g. the word ‘much’ in the question is not highlighted by CP-LRP, while AttnLRP highlights it stronger and AtMan focuses excessively on it. For comparison, we also visualize a random attribution with Gaussian noise.

ment should be studied in future work.

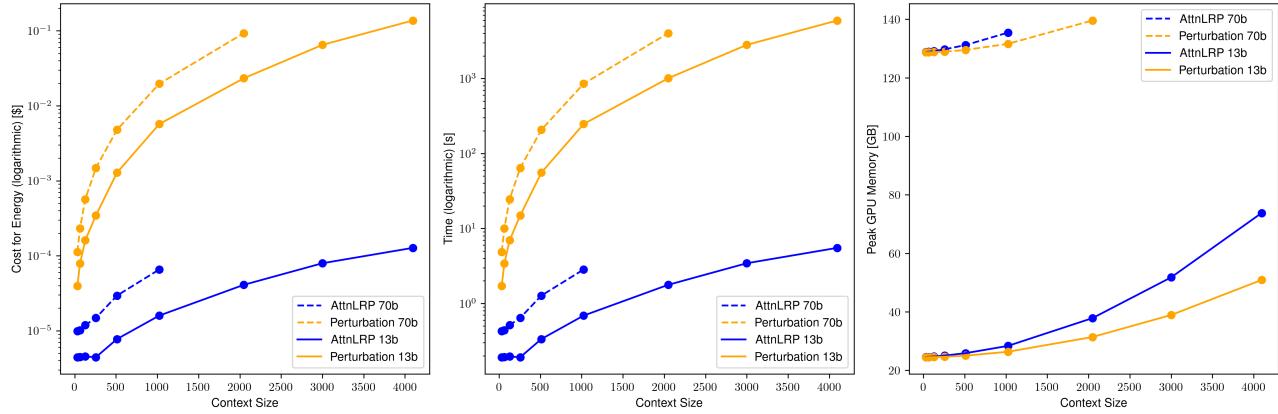


Figure 8. From left to right: Cost in dollar, time in seconds and peak GPU memory in gigabytes for AttnLRP and linear-time perturbation. Evaluated on Llama 2-70b and Llama 2-13b models on a node with four A100-SXM4 40GB. $G \times$ AttnRollout is in the range of AttnLRP and omitted for clarity of visualization. Because AttnLRP consumes more than 160 GB of RAM, the curves for the 70b model stop. Measured at fixed intervals of context size 32, 64, 128, 256, 512, 1024, 2048, 3000, 4096.

Table 5. ViT Perturbation Experiment (Faithfulness). For SmoothGrad, we set $\sigma = 0.01$, for AtMan $p = 1.0$ and $t = 0.1$, for AttnRoll $dt = 0.99$, and for $G \times$ AttnRoll $dt = 0.91$. “all epsilon” indicates that the ε -rule has been used on the linear and convolutional layers. The term “best” refers to the utilization of LRP with the composite proposed in B.4. ΔA^F denotes the area under the curve for a flipping perturbation experiment which leverages both A_{MoRF}^F of the most relevant first order, and A_{LeRF}^F of least relevant first order. ($\Delta A^F = A_{LeRF}^F - A_{MoRF}^F$). As discussed in Section B.2, this is equivalent to *insertion* perturbation.

Methods	ViT-B-16		
	(↑) ΔA^F	(↓) A_{MoRF}^F	(↑) A_{LeRF}^F
Random	0.01	4.71	4.71
I×G	0.90	2.78	3.69
IG	1.54	2.55	4.10
SmoothG	-0.04	3.63	3.58
GradCAM	0.27	5.35	5.63
AttnRoll	1.31	4.866	6.17
$G \times$ AttnRoll	2.60	4.01	6.22
AtMan	0.70	5.57	6.27
CP-LRP (all epsilon)	2.53	2.45	4.98
γ CP-LRP (best)	6.06	1.53	7.59
AttnLRP (all epsilon)	2.79	5.22	2.42
AttnLRP (best)	6.19	1.48	7.67

Table 6. Wikipedia Perturbation Experiment (Faithfulness). For SmoothGrad, we set $\sigma = 0.1$, for AtMan $p = 1.0$, for AttnRoll $dt = 1$, and for G×AttnRoll $dt = 1$. "all epsilon" indicates that the ε -rule has been used on all linear layers. ΔA^I denotes the area under the curve for the *insertion* perturbation experiment which leverages both A_{MoRF}^I of the most relevant first order, and A_{LeRF}^I of least relevant first order. ($\Delta A^I = A_{MoRF}^I - A_{LeRF}^I$). As discussed in Section B.2, this is equivalent to *flipping* perturbation.

Methods	Llama 2-7b		
	Wikipedia		
	(↑) ΔA^I	(↑) A_{MoRF}^I	(↓) A_{LeRF}^I
Random	-0.07	2.31	2.38
I×G	0.18	1.27	1.09
IG	4.05	3.74	-0.31
SmoothG	-2.22	0.68	2.90
GradCAM	2.01	2.36	0.35
AttnRoll	-3.49	1.46	4.95
G×AttnRoll	9.79	8.79	-1.00
AtMan	3.31	4.06	0.76
CP-LRP (all epsilon)	7.85	6.43	-1.42
AttnLRP (all epsilon)	10.93	9.08	-1.85

Table 7. IMDB Perturbation Experiment (Faithfulness), For SmoothGrad we set $\sigma = 0.05$, for AtMan $p = 0.7$, for AttnRoll $dt = 1$, and for G×AttnRoll $dt = 1$. "all epsilon" indicates that the ε -rule has been used to propagate relevance to the layers. ΔA^I demonstrates the area under the curve for the perturbation experiment of the type *Insertion* which leverages insights from both A_{MoRF}^I of the most relevant first order, and A_{LeRF}^I of least relevant first order. ($\Delta A^I = A_{MoRF}^I - A_{LeRF}^I$). As discussed in Section B.2, this is equivalent to *flipping* perturbation.

Methods	Llama 2-7b		
	IMDB		
	(↑) ΔA^I	(↑) A_{MoRF}^I	(↓) A_{LeRF}^I
Random	-0.01	-0.47	-0.46
I×G	0.12	-0.69	-0.81
IG	1.23	-0.06	-1.29
SmoothG	0.25	-0.74	-0.98
GradCAM	-0.82	-1.10	-0.28
AttnRoll	-0.64	-0.64	0.00
G×AttnRoll	1.61	0.77	-0.84
AtMan	-0.05	-0.54	-0.49
CP-LRP (all epsilon)	1.72	0.50	-1.22
AttnLRP (all epsilon)	2.50	1.12	-1.38

Table 8. SQuAD v2 Perturbation Experiment (Faithfulness). For SmoothGrad, we set $\sigma = 0.1$ and for AtMan $p = 0.9$, for AttnRoll $dt = 1$, and for G×AttnRoll $dt = 1$. "all epsilon" indicates that the ε -rule has been used on all linear layers. ΔA^I denotes the area under the curve for the *insertion* perturbation experiment which leverages both A_{MoRF}^I of the most relevant first order, and A_{LeRF}^I of least relevant first order. ($\Delta A^I = A_{MoRF}^I - A_{LeRF}^I$). As discussed in Section B.2, this is equivalent to *flipping* perturbation.

Methods	Flan-T5-XL		
	SQuAD v2		
(↑) ΔA^I	(↑) A_{MoRF}^I	(↓) A_{LeRF}^I	
Random	0.01	-6.05	-6.06
I×G	0.27	-5.51	-5.78
IG	0.77	-5.31	-6.08
SmoothG	0.16	-5.47	-5.63
GradCAM	0.94	-5.24	-6.19
AttnRoll	-0.42	-5.96	-5.54
G×AttnRoll	-0.06	-5.65	-5.60
AtMan	1.01	-5.18	-6.19
CP-LRP (all epsilon)	1.74	-4.93	-6.67
AttnLRP (all epsilon)	1.76	-4.91	-6.67

The San Diego Electric Railway (SDERy) was a mass transit system in Southern California, United States, using 600 volt DC streetcars and (in later years) buses.
Maintenance of way (commonly abbreviated to MOW) refers to the maintenance, construction, and improvement of rail infrastructure, including tracks, ballast, grade, and lineside infrastructure such as signals and signs.
France currently operates the second-largest European railway network, with a total of 29,901 kilometres of railway.
The MHR had, in 1846, amalgamated with the "Little" North Western Railway (NWR), which was taken over by the Midland Railway in 1874. Awdry, p.97. The rival London and North Western Railway (LNWR) built its own branch line to Morecambe in 1864, joining the main LNWR line at Hest Bank.
Some railway companies had a standard signalbox design, such as the London & North Western Railway, whereas others, such as the Great Eastern Railway had many different designs.
There are more than 16,000 student tour operators and travel agencies estimated in this market.
In 2010 it was totally integrated with the main regional public transport company, ARST (Azienda Regionale Sarda Trasporti).
A school was operating in the town in 1914.
The railway opened in 1886 with four stations using steam locomotives hauling unheated wooden carriages; in the next six years the line was extended and three more stations opened.
The Montreal Jonqui train (formerly the Saguenay) is a passenger train operated by Via Rail between Montreal and Saguenay (borough of Jonqui) in Quebec, Canada.

Figure 9. AttnLRP attributions on top 10 ActMax sentences collected over the Wikipedia summary dataset for neuron #256, in layer 18. The knowledge neuron seems to activate for transport systems (railways in particular).

The Court held unanimously in favor of a schoolteacher fired for her critical remarks in conversations with her principal.
The town schoolteacher was reading the book to her students when she was asked by her husband, the postmaster, to help name the little settlement.
She was a teacher for forty years and her writing has appeared in journals and anthologies since the early 1980s.
Her case made headlines and was covered by major news networks for being a notorious teacher who had an unlawful sexual relationship with one of her students.
His departure in 1971 generated some controversy on campus; he was regarded as an excellent teacher by his students, however, the administration was viewed as being more concerned about research than education when making its tenure decisions.
The volume presents six short stories, with the titular story featuring Yahiro, a substitute teacher, who begins having an affair with his student Kago.
In 2018, Derek Michael Boyce, a high school math and science teacher at the school, was arrested for having an inappropriate relationship with one of his students, a fifteen-year-old girl.
The film follows a school teacher as she suspects one of her students is suffering from personal problems in his home life, not knowing that the student is harboring an evil demon in his house.
During his time as a teacher Franco admitted to having sex with several of his students, which led to lawsuits and a \$2 million sexual-misconduct settlement in 2021.
It tells the story of a schoolteacher who falls in love with one of his students, and moves away in order to escape his infatuation.

Figure 10. AttnLRP attributions on top 10 ActMax sentences collected over the Wikipedia summary dataset for neuron #2207, in layer 20. The knowledge neuron is activating for 'teacher', in unusual context such as inappropriate behavior, sexual misconduct etc.

This contrasts with the pattern in all vascular plants (seed plants and pteridophytes), where the diploid sporophyte generation is dominant.
Like most other plants in the family, these produce umbels of flowers. Genus of the Month: Zizia.
Fructification () are the generative parts of the plant (flower and fruit) (as opposed to its vegetative parts, trunk, roots and leaves).
Symptoms of red stele can include a red core in the roots, wilting of leaves, reduced flowering, stunting, and bitter fruit.
Catopsis is a genus in the botanical family Bromeliaceae, subfamily Tillandsioideae.
Dracaena ciliensis, synonym Sansevieria ciliensis, is a xerophytic CAM succulent native to a small region of Somalia near the town of Eyl.
Dracaena suffruticosa, synonym Sansevieria suffruticosa is a species of Dracaena native to eastern Africa, from Ethiopia to Malawi.
Like other plant families, the Solanaceae is divided further into subfamilies, tribes and subtribes.
Pathogens that cause wilting diseases invade the vascular vessels and cause the xylem to fail to transport water to the foliage, thus causing wilting of stems and leaves.
Dracaena hammingtonii, synonym Sansevieria ehrenbergii, (blue sansevieria, sword sansevieria, oldupai, or East African wild sisal) is a flowering plant which grows in northeastern and eastern tropical Africa (Djibouti, Eritrea, Ethiopia, Kenya, Somalia, Sudan and Tanzania) and the Arabian Peninsula (Oman and Saudi Arabia).

Figure 11. AttnLRP attributions on top 10 ActMax sentences collected over the Wikipedia summary dataset for neuron #922, in layer 18. The knowledge neuron seems to be activating for scientific descriptions of plants.

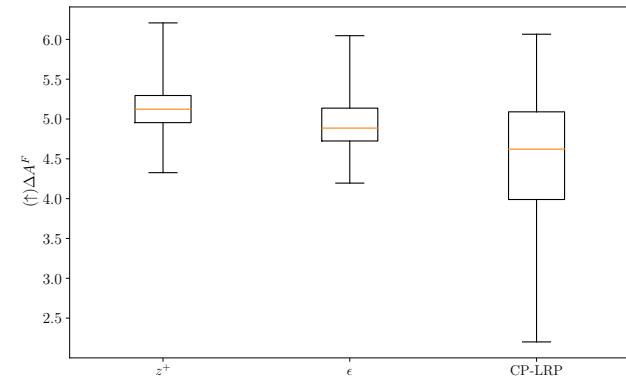


Figure 12. Statistics on Rules used for softmax layers: Either applying z^+ , ϵ -rule, or regarding as constant as proposed in CP-LRP. Propagating relevance values through (specifically by applying z^+ rule) softmax improves the faithfulness of explanations compared to the case where we block its propagation.

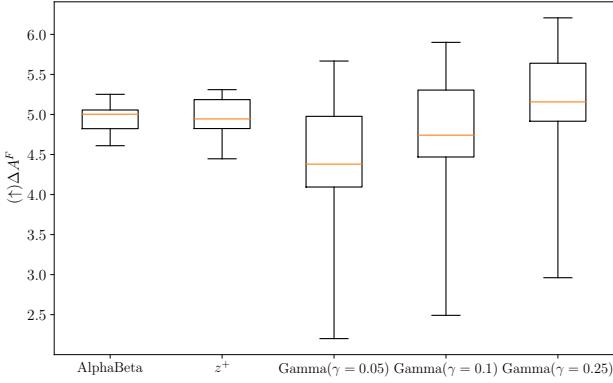


Figure 13. Statistics on Rules used for *Convolution* layers: Applying z^+ and AlphaBeta proposes acceptable results however the most faithful results can be reached via Gamma($\gamma = 0.25$).

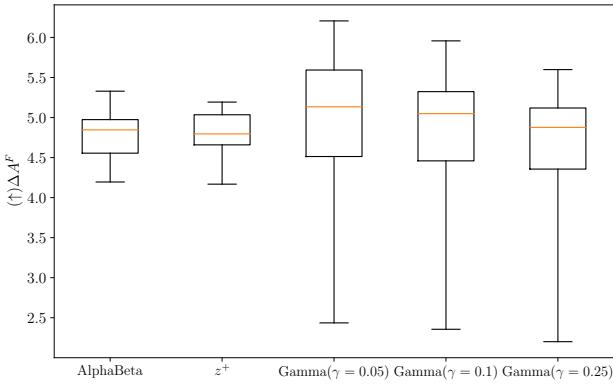


Figure 14. Statistics on Rules used for *Linear* layers: Similar to *Convolution* layers, Gamma seams more promising however with different γ value (0.05 in this case).

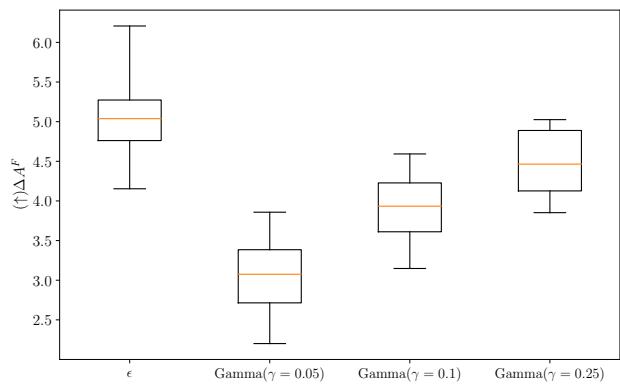


Figure 16. Statistics on Rules used for *LinearOutputProjection* layers: The ϵ -rule outperforms other rules clearly.

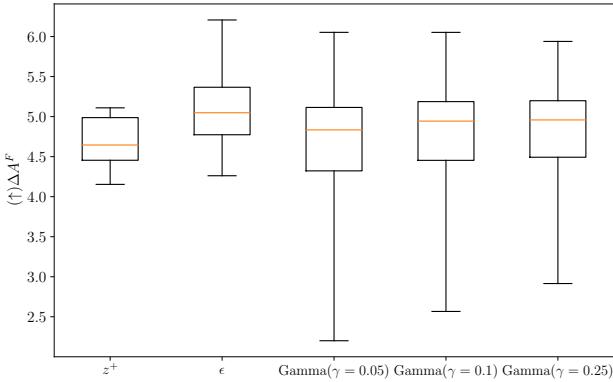


Figure 15. Statistics on Rules used for *LinearInputProjection* layers: Gamma and ϵ rules are competitive in this case, however since there is larger difference between the minimum and the lower quartile in Gamma rules, the most faithful choice will be ϵ -rule.