Model Interpretability and Rationale Extraction by Input Mask Optimization

Marc Brinner and **Sina Zarrieß**
Bielefeld University
Faculty for Linguistics and Literary Studies
{marc.brinner,sina.zarriess}@uni-bielefeld.de

Abstract

Concurrent with the rapid progress in neural network-based models in NLP, the need for creating explanations for the predictions of these black-box models has risen steadily. Yet, especially for complex inputs like texts or images, existing interpretability methods still struggle with deriving easily interpretable explanations that also accurately represent the basis for the model's decision. To this end, we propose a new, model-agnostic method to generate extractive explanations for predictions made by neural networks, that is based on masking parts of the input which the model does not consider to be indicative of the respective class. The masking is done using gradient-based optimization combined with a new regularization scheme that enforces sufficiency, comprehensiveness, and compactness of the generated explanation. Our method achieves state-of-the-art results in a challenging paragraph-level rationale extraction task, showing that this task can be performed without training a specialized model. We further apply our method to image inputs and obtain high-quality explanations for image classifications, which indicates that the objectives for optimizing explanation masks in text generalize to inputs of other modalities.

1 Introduction

Black-box machine-learning models like transformers (Vaswani et al., 2017) or convolutional neural networks (Tan and Le, 2019) are state-of-the-art in natural language processing and computer vision. Their complexity enables them to perform well on a variety of tasks, but this comes at the cost of a lack of interpretability: The question of why a model made a specific prediction cannot be answered reliably. Especially if such black-box models are used in critical real-world applications (e.g., in the medical domain), this creates a demand for methods that explain network predictions while fulfilling a variety of requirements, like being easy to implement, model and task agnostic, faithful to the inner workings of the network, and producing results that are easily interpretable for humans.

To this end, a variety of interpretability methods have been proposed (Guidotti et al., 2018; Zhang et al., 2021), but as the aforementioned requirements are often at odds, at least one of them often remains unfulfilled. Reasons for this include the reliance on complex message passing schemes that require laborious implementations (e.g., Montavon et al., 2017, Shrikumar et al., 2017), the applicability only to specific model architectures (e.g., Yuan et al., 2021, Abnar and Zuidema, 2020), or the fact that explanations often highlight individual, disconnected input features (e.g., standard gradient-based saliency), which contradicts human intuition of a sensible explanation (compare Section 2.1 for details).

As an example, in a text classification setting, interpretability methods often highlight individual words that explain the prediction, but do not include their context (Remmer, 2022), even though the context of a word is crucial in determining its meaning: The word "good" influences the prediction in a completely different way if it is preceded by the word "not", meaning that this context has an impact on the classification and should therefore be part of the rationale. Notably, this holds true even in the absence of such modifiers, since the context must be available to confirm this absence.

In this work, we propose a new method for model explainability that is able to identify parts of the input that are, on the one hand, most indicative of a class and, on the other hand, perceived as a sensible rationale by humans. Our method is applicable to all input types that define a spatial structure between individual features (e.g., texts, images) and builds on the assumption that interpretable explanations correspond to smooth and connected regions of features with respect to this spatial structure. It uses numerical optimization to mask out parts of the input that the model does not consider

Findings of the Association for Computational Linguistics: ACL 2023, pages 13722–13744
July 9-14, 2023 ©2023 Association for Computational Linguistics

indicative of the class of interest, thus leaving only the parts of the input that are indicative of this class. The masking is done using gradient-based optimization combined with a new regularization scheme that enforces sufficiency, comprehensiveness, and compactness of the generated explanation (Yu et al., 2019), three criteria that have been established in the domain of rationale extraction but are less common in network interpretability methods. In this way, our method bridges the gap between model interpretability and rationale extraction, thereby showing that the latter of which can be performed without training a specialized model, only on the basis of a trained classifier.

2 Background

Methods that explain the predictions made by black-box models to users can be broadly categorized into (i) interpretability methods that aim at creating explanations for existing classifiers after they have been trained (Section 2.1) and (ii) rationale extraction approaches that are designed to create a rationale as a model output in addition to the usual label prediction (Section 2.2). Our interpretability method relies on gradient-based input optimization, discussed in detail in Section 2.3.

2.1 Neural Network Interpretability

Interpretability methods usually assign importance scores to features or parts of a given input, indicating how relevant the respective feature is for making the prediction. Many early methods focus on convolutional neural networks and use backpropagation-like procedures to compute saliency scores for each input feature. Simonyan et al. (2013) use the network's gradient at the input image as saliency scores, while Sundararajan et al. (2017)'s integrated gradients method sums over gradients at different inputs that are created by gradually transforming a neutral input into the input of interest. The DeconvNet architecture (Zeiler and Fergus, 2014) and the guided backpropagation algorithm (Springenberg et al., 2015) again rely on a single evaluation but change the standard gradient computation to produce visually improved importance maps. Attribution methods like layer-wise relevance propagation Bach et al. (2015) extend this idea by defining a backward pass that redistributes the total function value layer-wise backwards using a propagation rule that makes the total relevancy within each layer add up to the function value that is to be explained. Deep Taylor Decomposition (Montavon et al., 2017) and DeepLIFT (Shrikumar et al., 2017) then introduced different rules for redistributing the relevance between layers. For transformer models, methods like (Abnar and Zuidema, 2020) track the attention flow through the network. This has been extended to incorporate information from attribution methods like the Deep Taylor Decomposition to more accurately identify neurons that have a strong influence on the final prediction (Chefer et al., 2021b,a).

Other well-known explainability methods rely on input perturbations. LIME (Ribeiro et al., 2016) identifies important input features by perturbing the input, observing the change in the model predictions, and fitting an interpretable model to the observed data, while other methods occlude parts of the input to detect features that are important for the classification (Zeiler and Fergus, 2014; Bazzani et al., 2016; Zhou et al., 2014; Petsiuk et al., 2018).

A further approach to model interpretability is to generate an input that maximally activates specific neurons, thereby yielding insights about the responsibilities of these neurons, as was done for CNNs by Simonyan et al. (2013). Fong and Vedaldi (2017) then used a similar idea to remove class-indicative information from input images to detect the parts of the image responsible for the classification.

The method we propose in this paper differs from standard gradient-based techniques by not relying on evaluations at a single point or at fixed perturbations, but at points that are determined by a dynamic optimization process. This control via optimization is also a key difference from methods that rely on random permutations or masking of input features. Compared to message-passing schemes and model-specific methods (like methods for transformer interpretability), relying only on the gradient makes our method applicable to models with a variety of architectures and layer types without requiring additional implementational effort.

2.2 Rationale Extraction

The task of *rationale extraction*, also commonly referred to as *selective rationalization*, is concerned with designing models that can produce human-interpretable rationales in addition to the usual model output (Lei et al., 2016), with the domain usually being textual inputs and the rationales being a subset of the input text that is determined to be responsible for the prediction. Lei et al.

(2016) approached this task by developing a two-step procedure in which a proposal network extracts a rationale from the input text and a subsequent classification network only has access to the rationale to make the final prediction. By training this model end-to-end, the proposal network learns to extract the most useful text fragments from the input, which thus corresponds to an explanation for the classification. Later, Yu et al. (2019) proposed three criteria that rationales should satisfy to be perceived as sensible:

Sufficiency: The rationale should be sufficient to correctly classify the sample only by its rationale.

Comprehensiveness: All relevant information should be contained in the rationale, meaning that the correct label can not be inferred by just considering the words not included in the rationale.

Compactness: The rationale should be sparse but should nevertheless consist of consecutive text fragments instead of single words.

Yu et al. (2019)'s methods enforce these criteria through regularizers and by using a complement predictor that predicts the correct label based on all words that are not part of the rationale. Training the proposal network to fool the complement predictor then enforces the comprehensiveness constraint. Other approaches extend this and extract class-dependent rationales (Chang et al., 2019) or select complete paragraphs as rationales (Chalkidis et al., 2021).

The two main differentiators of rationale extraction models to the interpretability methods discussed in Section 2.1 are that, one the one hand, models are explicitly trained to produce rationales instead of creating them post hoc, and, on the other hand, the focus is on creating human interpretable rationales while the focus for interpretability methods often is on mathematical faithfulness measures.

Our method combines the focus on faithfulness with the desire for human interpretability to create rationales that faithfully explain model predictions post hoc *and* correspond to human rationales, as these properties substantially enhance the usefulness of explanations for many applications.

2.3 Input Optimization

As mentioned in Section 2.2, optimization of input images for CNNs has been used to explain the responsibilities of specific neurons, but notably, the resulting images do not resemble naturally occurring images. This is caused by the huge complexity and highly nonlinear behavior of neural networks, leading to the property of having unpredictable behavior on out-of-domain inputs that quickly arise during the optimization. In different experiments, this has led to behaviors like making highly confident class predictions for images that resemble random noise (Nguyen et al., 2015) or predicting a completely different class after adding almost imperceivable noise to a given image (Szegedy et al., 2014). Unconstrained optimization of the input to a neural network to optimize the activation of specific neurons will therefore inevitably result in inputs that are out-of-domain, do not resemble natural images, or seem downright counter-intuitive. Different strategies for mitigating this problem in the context of input optimization exist (e.g., the use of GANs, Nguyen et al., 2016), with the most common being extensive regularization to prevent high-frequency information in images from influencing the prediction (Yosinski et al., 2015; Mahendran and Vedaldi, 2015) or using lower-resolution inputs and blurring to limit the degrees of freedom within the optimization (Fong and Vedaldi, 2017).

In this study, we use input optimization to perform model interpretability by optimizing a mask to suppress all parts of a given input that a given model does not consider indicative of the given class. Compared to (Fong and Vedaldi, 2017), we propose a new optimization objective as well as a new regularization scheme that allows for the creation of more detailed masks. Additionally, we expand the scope of input optimization methods from the domain of images to text processing.

3 MaRC

In this section, we introduce **MaRC**, our framework for **Ma**sk-based **R**ationale **C**reation. Section 3.1 develops the general framework. Sections 3.2 and 3.3 address the specificities of applying MaRC to texts and images, respectively.

3.1 Method

We design an interpretability method that detects parts of an input x that a model M considers most indicative of a specific class c. We assume an input x with n input features, each of which could be high-dimensional, e.g., token embeddings or pixels with color channels. The main idea of the approach is to detect input features that are highly indicative of class c by replacing as much of the input as possible with an uninformative input b, i.e., an input that the model does not consider indicative of any class, while having the model assign a high score for class c to the altered input. We define a mask $\lambda \in \mathbb{R}^n$, $\lambda_i \in [0, 1]$ to obtain a masked input $\tilde{x}$ in the following way:

$$\tilde{x} = \lambda \cdot x + (1 - \lambda) \cdot b \tag{1}$$

When λ_i is close to 1, feature i is mostly retained in $\tilde{x}$, while λ_i values close to 0 replace feature i almost completely with the uninformative b.

MaRC tackles this masking as an optimization problem: it optimizes λ to obtain rationales that fulfill the properties of sufficiency, comprehensiveness, and compactness (compare Section 2.2). It models these properties via dedicated regularizers, which we will develop step by step in the following.

Sufficiency We want to find a mask λ such that the probability that model M assigns to $\tilde{x}$ for class c is close to 1. We optimize this criterion as follows:

$$\underset{\lambda \in [0,1]^n}{\arg\min} \quad - \mathcal{L}(\tilde{x}, c) + \underbrace{\alpha_\lambda \left[\frac{1}{n} \sum_{i=1}^{n} \lambda_i \right]^2}_{\Omega_\lambda} \tag{2}$$

Here, $\mathcal{L}(\tilde{x}, c)$ is a scoring function for c under M and Ω_λ is a sparsity regularizer that enforces the detection of the smallest set of input features that still induces a high score for c. An obvious choice for $\mathcal{L}(\tilde{x}, c)$ is the log-likelihood of c, maximizing the probability of c under M and leading λ to highlight *class-discriminative information*, i.e., input features that indicate *only* class c. A different choice would be the logarithm of the sigmoid of the logit for c, which does not suppress other classes and therefore leads λ to highlight *class-indicative information*, i.e., all input features relevant for c, even if they are indicative of other classes as well. In both cases, M considers $\tilde{x}$ to be highly indicative of class c, thereby fulfilling the sufficiency criterion.

Comprehensiveness Optimizing Equation 2 leads to sufficiency but not comprehensiveness, as the smallest set of highly indicative input features is detected. To detect *all* information relevant for c, we introduce the *complement of rationale* (Yu et al., 2019):

$$\tilde{x}^{\mathsf{c}} = (1 - \lambda) \cdot x + \lambda \cdot b \tag{3}$$

which leaves features unmasked that were masked for $\tilde{x}$. Minimizing the score of $\tilde{x}^{\mathsf{c}}$ for c enforces all parts that indicate class c to be masked in $\tilde{x}^{\mathsf{c}}$ (meaning that they will be unmasked in $\tilde{x}$), resulting in the following optimization:

$$\underset{\lambda \in [0,1]^n}{\arg\min} \quad - \mathcal{L}(\tilde{x}, c) + \mathcal{L}(\tilde{x}^{\mathsf{c}}, c) + \Omega_\lambda \tag{4}$$

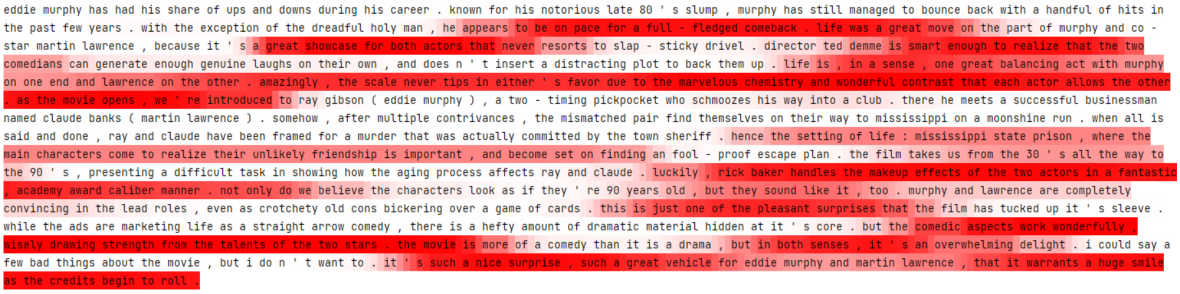
This formulation combines the "deletion game" and "preservation game" that were introduced by Fong and Vedaldi (2017) but treated as separate objectives. Optimizing the mask with respect to both objectives greatly supports the detection of precise boundaries of the relevant features.

Compactness The original compactness criterion states that a rationale shall consist of longer but fewer meaningful spans of *text*. Here, we generalize this to all input types that possess a spatial structure that defines neighborhoods around input variables. The underlying assumption is, that for these types of inputs, a feature is only meaningful in the context of its neighborhood, as, for example, is the case for single words in text or individual pixels in images, so that a sensible rationale must include larger groups of closely located features.

Thus, we now assume a general spatial structure on the input x that defines distances $d(i, j)$ between the features i and j, with features that are closer together having a higher chance of belonging to the same meaningful entity. We enforce the selection of larger groups of features by reparameterizing our mask, i.e., we introduce two new parameters, $w \in \mathbb{R}^n$ and $\sigma \in \mathbb{R}^n_{>0}$ from which the mask values λ can subsequently be calculated. The optimization is then performed with respect to w and σ.

The mask values λ are mainly determined by w, in a way that w_i largely determines the final value of λ_i. Crucially, w_i now also influences the values of λ around i, so that, for example, λ_{i-1} and λ_{i+1} are also strongly influenced by w_i. σ_i then determines the strength and extent of w_i's influence on its neighbors, as it parameterizes an unnormalized Gaussian placed at position i, so that the influence $w_{i \to j}$ of a weight w_i onto λ_j is then given by:

$$w_{i \to j} = w_i \cdot \exp\big(- \frac{d(i, j)^2}{\sigma_i} \big) \tag{5}$$



eddie murphy has had his share of ups and downs during his career . known for his notorious late 80 ' s slump , murphy has still managed to bounce back with a handful of hits in the past few years . with the exception of the dreadful holy man , he appears to be on pace for a full - fledged comeback . life was a great move on the part of murphy and co - star martin lawrence , because it ' s a great showcase for both actors that never resorts to slap - sticky drivel . director ted demme is smart enough to realize that the two comedians can generate enough genuine laughs on their own , and does n ' t insert a distracting plot to back them up . life is , in a sense , one great balancing act with murphy on one end and lawrence on the other . amazingly , the scale never tips in either ' s favor due to the marvelous chemistry and wonderful contrast that each actor allows the other . as the movie opens , we ' re introduced to ray gibson (eddie murphy) , a two - timing pickpocket who schmoozes his way into a club . there he meets a successful businessman named claude banks (martin lawrence) . somehow , after multiple contrivances , the mismatched pair find themselves on their way to mississippi on a moonshine run . when all is said and done , ray and claude have been framed for a murder that was actually committed by the town sheriff . hence the setting of life : mississippi state prison , where the main characters come to realize their unlikely friendship is important , and become set on finding an fool - proof escape plan . the film takes us from the 30 ' s all the way to the 90 ' s , presenting a difficult task in showing how the aging process affects ray and claude . luckily , rick baker handles the makeup effects of the two actors in a fantastic , academy award caliber manner . not only do we believe the characters look as if they ' re 90 years old , but they sound like it , too . murphy and lawrence are completely convincing in the lead roles , even as crotchety old cons bickering over a game of cards . this is just one of the pleasant surprises that the film has tucked up it ' s sleeve . while the ads are marketing life as a straight arrow comedy , there is a hefty amount of dramatic material hidden at it ' s core . but the comedic aspects work wonderfully , wisely drawing strength from the talents of the two stars . the movie is more of a comedy than it is a drama , but in both senses , it ' s an overwhelming delight . i could say a few bad things about the movie , but i do n ' t want to . it ' s such a nice surprise , such a great vehicle for eddie murphy and martin lawrence , that it warrants a huge smile as the credits begin to roll .

Figure 1: An exemplary rationale created by MaRC for the prediction of the *positive* sentiment label.

The final value for λ_j is then calculated as follows:

$$\lambda_j = \text{sigmoid}(\sum_i w_{i \to j}) \tag{6}$$

This parameterization of λ enforces neighboring inputs to have similar values if the corresponding σ values are large, which also plays a key role in regularizing the optimization to avoid the issues discussed in Section 2.3. Large values for σ are softly enforced by introducing an additional regularizer:

$$\Omega_\sigma = -\alpha_\sigma \cdot \frac{1}{n} \sum_{i=1}^{n} \log(\sigma_i) \tag{7}$$

The logarithm was chosen to enforce positive values of σ_i while gradually discounting the effect that increases in σ_i have on the loss function. Notably, this regularizer does not enforce large values of σ by means of hard constraints, meaning that low values and therefore sharper boundaries between mask values for neighboring features can be optimal if the other parts of the optimization objective support this behavior. This is in contrast to (Fong and Vedaldi, 2017), who used a lower resolution mask in combination with upsampling and Gaussian blur to detect smooth masks, which does not allow for sharp masks even if they were optimal.

In summary, the final optimization objective looks as follows:

$$\underset{w,\sigma \in \mathbb{R}^n}{\arg\min} \quad -\mathcal{L}(\tilde{x}, c) + \mathcal{L}(\tilde{x}^\mathsf{c}, c) + \Omega_\lambda + \Omega_\sigma \tag{8}$$

This objective can be optimized using stochastic gradient descent, but in practice, we found using an optimizer that incorporates momentum (e.g., Adam, Kingma and Ba, 2015) to be key for avoiding local optima and obtaining optimal results.

3.2 Textual Inputs

As MaRC only requires the gradient of a model prediction at the input, it can be applied to all common text processing models. In the following, we discuss specific aspects of using MaRC with state-of-the-art transformer architectures like BERT (Devlin et al., 2019).

As uninformative input b, we choose a sequence of *PAD*-tokens of the same length as x. During training, the model learns to treat these tokens as uninformative since they are added to inputs irrespective of their content or the desired output.

As we want importance scores for each individual word, we define n to be the number of words in the input sequence. Notably, this is different from the actual input dimension, as it is common to use WordPiece embeddings (Wu et al., 2016) which could split words into multiple input tokens. In this case, we use parameter tying to only have a single parameter for all pieces of a word representation. The distance function is then simply defined as $d(i, j) = |i - j|$, with i and j being the positions of the words in the text.

Finally, we found that introducing noise into the optimization process is beneficial for regularization (see Section 2.3 for discussion of regularization in input optimization). Thus, for text inputs, we add Gaussian noise to $\tilde{x}$ and $\tilde{x}^\mathsf{c}$ and randomly set mask values to 0 or to 1 in each optimization step.

3.3 Image Inputs

Image inputs also fulfill the requirements on the presence of a spatial structure that is needed for our method. They also provide natural choices for uninformative inputs, as uniformly colored images can generally be assumed to be uninformative in most prediction settings. Therefore, obvious choices for b would, for example, be a white image, a black image, or an image of the mean color within the given dataset. A different option is to remove usable information from the input image by blurring it and using this blurred image as uninformative input (Fong and Vedaldi, 2017). As parts of the input image could have the same color as the uninforma-

Method	Token F1	mAP	IoU F1	Suff.↓	Comp.↑
MaRC	**.473**	**.469**	**.163**	.028	.518
Occlusion	.432	.448	.125	.022	.415
Saliency$_n$	.435	.392	.04	.132	.287
Saliency$_s$	.425	.340	.076	.260	.246
InXGrad$_n$	.436	.396	.040	.136	.292
InXGrad$_s$	.425	.340	.084	.239	.248
Int. Grads$_n$	.428	.369	.036	.122	.274
Int. Grads$_s$	.431	.381	.071	.048	.528
LIME	.436	.380	.076	.047	.496
Shapley	.428	.439	.079	**-.015**	**.728**
Noise	.454	.450	.139	.034	.487
Ω_λ	.349	.350	.046	.120	.266
Ω_σ	.447	.425	.091	.036	.535
$\mathcal{L}(\tilde{x}^{\mathsf{c}}, c)$	.396	.436	.123	.052	.304

Table 1: Results on rationale extraction on the movie reviews dataset (DeYoung et al., 2020), including faithfulness evaluation. See Section A for an overview of the methods tested and for experimental details.

	ResNet-101		ViT-B/16	
Method	Suff. ↓	Comp.↑	Suff. ↓	Comp.↑
MaRC	**.196**	.612	**.139**	.596
M-Perturb	.260	.605	.174	.572
Grad-CAM	.197	.600	.161	.640
Exc-BP	.302	.600	-	-
Saliency	.442	.599	.355	.528
InputXGrad	.430	.586	.366	.506
Guided BP	.343	.630	-	-
Intgr. Grads	.344	**.641**	.261	.641
Occlusion	.324	.606	.194	.486
Attention	-	-	.241	.562
Attribution	-	-	.176	.608
Rollout	-	-	.205	.580
TAM	-	-	.146	**.658**

Table 2: Results for the faithfulness evaluation of different explainability methods for ResNet-101 and ViT-B/16. Compare Section A for an overview of the methods tested and for experimental details.

tive input (which renders the corresponding mask values meaningless) and even uniformly colored patches could be seen as informative by neural networks, we chose to alter the optimization objective to be the average over different choices for b, with B being the set of all uninformative inputs:

$$\underset{w,\sigma\in\mathbb{R}^{w\times h}}{\arg\min} \; \frac{1}{|B|}\sum_{b\in B}\mathcal{L}(\tilde{x}(b,\tau),c) - \mathcal{L}(\tilde{x}^{\mathsf{c}}(b,\tau),c) + \Omega_\lambda + \Omega_\sigma + \Omega_{\text{NB}} \qquad (9)$$

As images generally have more variables and therefore more degrees of freedom in the optimization, further regularization is needed to obtain sensible optimization results. To this end, this formulation includes an additional regularizer Ω_{NB}, which denotes the average squared difference between mask values that are neighboring with respect to the 8-connected grid structure of the image, weighted by a corresponding parameter α_{NB}.

To complete the specification of the optimization problem, we define the distance function d between two pixels to be the euclidean distance between their two-dimensional position vectors in the image grid. In contrast to the textual inputs, the introduction of noise to the optimization process did not prove to be beneficial.

4 Experiments on Rationale Extraction

We evaluate MaRC on rationale extraction, a task that is concerned with predicting the correct label for a given textual input while also providing a subset of the input as a rationale for the prediction.

4.1 Data

We use the movie review data set (Zaidan et al., 2007) with 2000 movie reviews annotated with sentiment labels (*positive* or *negative*) as well as span-level rationales. We test on the additional rationales created by DeYoung et al. (2020), which are more comprehensive and thus, on average, comprise a much larger fraction of words (7.2% vs. 31.4%). As our approach is designed for extracting span-level rationales and most other datasets for rationale extraction are not annotated on span-level (DeYoung et al., 2020), this is the only dataset suitable for an evaluation of MaRC. We use a standard BERT$_{base}$ model (Devlin et al., 2019) and train it as a standard binary classification model on the training data, therefore only using the class labels and not the annotated rationales.

4.2 Evaluation

There are two common ways of evaluating the rationales produced by different models (DeYoung et al., 2020; Atanasova et al., 2020):

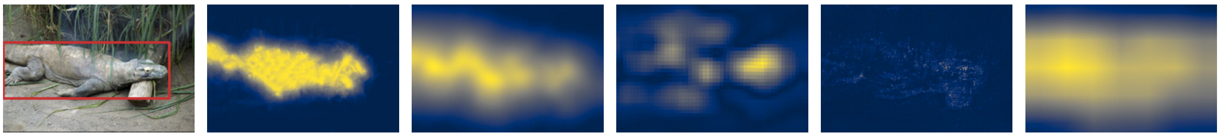

(a) Input (b) MaRC (c) M-Perturb (d) Occlusion (e) Integr. Grads (f) Grad-CAM

Figure 2: Comparison of masks created for ResNet-101 by different explainability methods.

1. **Agreement with human annotator rationales:** A strong overlap between rationales given by human annotators and rationales produced by an explainability model is a good indicator that sensible rationales have been selected. Additionally, similarity to human rationales could be considered a desirable property (depending on the use case), even if it is not perfectly in line with the actual reasoning process of the neural network.

2. **Faithfulness:** Ideally, the rationales produced by a model fulfill the conditions of sufficiency and comprehensiveness, meaning that they actually reveal the information that the model considered indicative for the predicted label.

Different metrics exist to evaluate the performance of approaches that produce "soft" scores (i.e., continuous values) or binary values as the output of the rationale generation. As we see use cases for both outputs, we evaluate our approach with respect to both. To create a binary mask from the continuous mask values that MaRC produces, we train a kernel regression model to predict the optimal percentage of words that need to be included in the rationale (described in Appendix A), which we do in the same way for all methods tested in this study.

To evaluate the agreement with human rationales, we calculate the token F1 score for the binary masks by using precision and recall of the "positive" class of words belonging to the rationale, while the soft-scoring models are evaluated using the mean average precision (mAP). To evaluate the agreement of larger detected spans with the spans present in the human rationales, we evaluate the IoU F1 score that counts a ground-truth span as correctly detected if there is a predicted span with an IoU of over 0.5, which again allows for the calculation of an F1 score for the "positive" class of detecting the spans. These three metrics were used in the ERASER benchmark (DeYoung et al., 2020), which also proposed metrics to evaluate sufficiency and comprehensiveness. For these metrics, we slightly deviate from their evaluation metrics by evaluating these scores for a given sample x and rationale r in the following way:

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

$$\text{sufficiency}(x, r) = \frac{1}{19} \sum_{i=1}^{19} M(x) - M(r_i) \quad (11)$$

Here, $M(x)$ denotes the class probability prediction (for the ground-truth class) of our model, r_i denotes the top $(i \cdot 5)\%$ of words according to the soft rationale scores (all other words are removed), and $x \backslash r_i$ denotes sample x with all words that belong to r_i removed, where we "remove" words by replacing the corresponding tokens with *PAD*-tokens. Therefore, the comprehensiveness score evaluates, how much removing the rationale decreases the model performance (higher scores are better) while the sufficiency score evaluates how well the correct label can be predicted from the rationale alone (lower scores are better).

4.3 Results

The evaluation results are displayed in the upper part of Table 1 (see Appendix A for more details on the setup). We compare MaRC against other interpretability methods that are commonly used in the context of NLP but omit specialized rationale extraction models as they (i) usually produce binary masks, making it impossible to perform the soft-scoring evaluation, and (ii) do not produce explanations for existing models, which makes the faithfulness evaluation inapplicable.

We see that MaRC achieves state-of-the-art results on all measures that evaluate agreement with human rationales, i.e. Token F1, mAP, and IoU F1, showing that MaRC is the best method for obtaining rationales that match human intuition. Especially with respect to the IoU F1 score, MaRC outperforms all other methods by a large margin, even though the hyperparameters for other methods were set to explicitly support high scores in this measure (e.g., masking larger spans for occlusion

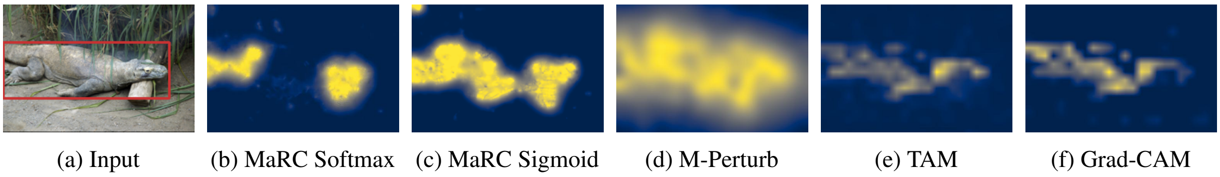

(a) Input (b) MaRC Softmax (c) MaRC Sigmoid (d) M-Perturb (e) TAM (f) Grad-CAM

Figure 3: Comparison of masks created for ViT-B/16 by different explainability methods.

and LIME). This highlights that MaRC is suitable for detecting span-level rationales in a paragraph-long text that agree with spans that humans annotate, without being trained to do so and without additional model components as in state-of-the-art rationale extraction models.

For sufficiency and comprehensiveness, MaRC also achieves impressive results, being outperformed in both metrics only by Shapley value sampling. The excellent performance of this method with regard to these evaluation metrics is not surprising, though, as it is based on choosing a random permutation of input features, adding them successively to the input, and using the change in the model's output as the resulting score. This method is very closely connected to the sufficiency and comprehensiveness calculations, thereby rendering the great results of this method unsurprising. It should be noted, that multiple methods, including MaRC, achieve close to optimal results for sufficiency, as scores close to 0 indicate that the removal of very few high-scoring tokens is enough to completely throw off the classifier. Notably, MaRC can produce good results while aiming to create human-like rationales, showing that this kind of rationale to some extent corresponds to the inner workings of the neural network.

We also conduct an ablation study that tests the importance of the different parts of the optimization objective by leaving these parts out in turn and reporting the results with the altered objective. The results are displayed in the lower part of Table 1, with the "Method" column specifying with part of the objective is omitted. The full optimization objective is almost uniformly the best-performing variant, proving that all parts a essential to achieve optimal performance.

5 Experiments on Image Classification

We evaluate MaRC on the task of creating rationales for classifications of ImageNet (Russakovsky et al., 2015) images. A visual comparison of masks created by MaRC and other interpretability approaches for ResNet-101 (He et al., 2016) and the vision transformer ViT-B/16 (Kolesnikov et al., 2021) is displayed in Figures 2 and 3, respectively (see Appendix C for with more visualizations). We see that MaRC is able to produce sharp masks that often cover the complete object of interest in the image. For ViT-B/16, we include a visualization highlighting the distinction between class-discriminative vs. class-indicative information (compare Section 3.1): Figure 3b) used the softmax of the model output as scoring function, which leads MaRC to highlight only the head and tail of the animal, the two parts that the model uses to differentiate the correct class from the other classes. For Figure 3c), on the other hand, a mixture of the sigmoid of the class logit and the softmax of the model output was used with a ratio of 9:1, making the model highlight all parts in the image that indicate the ground-truth class, as long as they are not significant indicators of other classes.

We also evaluate the faithfulness of the explanation created by MaRC and a variety of other interpretability methods using the same metrics as for textual inputs. For this experiment, we use pretrained ResNet-101 and ViT-B/16 models on a random sample of 500 ImageNet validation images, with further implementational details being described in Appendix A. As shown in Table 2, MaRC is the best-performing model with respect to sufficiency for both ResNet-101 and ViT-B/16, showing that the areas that MaRC highlights are indeed the areas that allow the model to predict the correct class based solely on these regions. With respect to comprehensiveness, MaRC achieves competitive results, only falling behind model-specific architectures that heavily use the knowledge about the inner workings of the model and the information flow inside it (e.g., transition attention maps (TAM), Yuan et al., 2021), as well as two other methods in the form of Guided Backpropagation (Springenberg et al., 2015) and Integrated Gradients Sundararajan et al. (2017). The latter two methods often predict individual pixels that are spread over many areas of the image as the most indicative input features, indicating that the removal of key pixels at different positions of the image is a good strategy to quickly decrease the classifier performance, an approach that MaRC is actively discouraged to pursuit.

6 Conclusion

We propose a new method for creating explanations for neural network predictions that are faithful to the model's reasoning process as well as being sensible with respect to human judgment. We achieve state-of-the-art results on the task of rationale extraction, achieve competitive or state-of-the-art results with respect to faithfulness, and provide visually sensible explanations for classifications of images. As MaRC is model-agnostic, we believe it to be a useful tool in many areas of machine learning that include textual or image inputs. We further believe that other domains can make use of MaRC, including multimodal tasks that, for example, combine textual and image inputs, as well as other domains that fulfill the requirements on the spatial structure of the input, like auditory data.

7 Limitations

Compared to other interpretability methods, MaRC is able to create explanations that more closely resemble human rationales. Nevertheless, the similarity to human rationales is always limited by the inner workings of the respective neural network: If a network's reasoning does not mirror human reasoning, the resulting rationales will be incomprehensible to humans.

Additionally, rationales created by MaRC are the result of a complete input optimization process. Therefore, the rationale creation usually requires hundreds of forward passes and gradient evaluations for the respective neural network, which makes the process of creating the rationale time-consuming and therefore infeasible for many real-time applications. On modern hardware, creating a rationale for BERT$_{base}$ can take two to three minutes depending on the length of the input text, while ResNet-101 and ViT-B/16 are faster at about one minute.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

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

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

Loris Bazzani, Alessandra Bergamo, Dragomir Anguelov, and Lorenzo Torresani. 2016. Self-taught object localization with deep networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9.

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

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

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

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

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

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

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

Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Jacob Gildenblat and contributors. 2021. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

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

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

Aravindh Mahendran and Andrea Vedaldi. 2015. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120:233–255.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436.

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

Eliott Remmer. 2022. Explainability methods for transformer-based artificial neural networks:: a comparative analysis.

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

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

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

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

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. In *Deep Learning Workshop, International Conference on Machine Learning (ICML)*.

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

Tingyi Yuan, Xuhong Li, Haoyi Xiong, Hui Cao, and Dejing Dou. 2021. Explaining information flow inside vision transformers using markov chain. In *eXplainable AI approaches for debugging and diagnosis*.

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

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

Jianming Zhang, Zhe L. Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2016. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126:1084–1102.

Yu Zhang, Peter Tino, Ales Leonardis, and Ke Tang. 2021. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742.

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

A Experimental Details

The implementations of MaRC for the experiments conducted in this study is available at `https://github.com/inas-argumentation/Explainability`.

A.1 Rationale Extraction

We perform rationale detection using BERT$_{base}$ (uncased) (Devlin et al., 2019), which we train as a binary classifier for at most 20 epochs on the first eight folds of the movie review dataset (Zaidan et al., 2007), with the ninth and tenth fold being used for validation and testing, respectively. For the optimization, we use the Adam optimizer and achieve a 96.5% test set accuracy.

For rationale creation, the results from Table 1 as well as the example images for MaRC were created by using the optimization objective given by Equation 8 with all specifications as described in Section 3.2, hyperparameters set to $\alpha_\lambda = 1$, $\alpha_\sigma = 1.2$ and w and σ being uniformly initialized to 1.2 and 2, respectively. We add zero-mean Gaussian noise to $\tilde{x}$ and $\tilde{x}^c$ ($\sigma = 0.03$) and randomly set 5% of mask values to 0 or 1, respectively, in each optimization step. We use the log-likelihood of the respective class as scoring function. All these choices were made by using the validation split, with the measure of quality being visual coherence of the created explanation, as the data set does not offer a validation split with the same label distribution, thus making validation with respect to scores infeasible. Texts that surpass the limit of 510 input tokens for BERT$_{base}$ are split into multiple segments, with consecutive segments overlapping for 100 tokens, and a separate mask is predicted for each segment. The resulting masks are concatenated, with the overlapping parts being linearly blended. We proceed in the same way for all other interpretability methods.

The following models and parameters were used in the method comparison:

- *Occlusion* (Zeiler and Fergus, 2014): We chose to mask slightly larger spans of 5 tokens as this produced smoother masks which resulted in higher IoU F1 scores. Occluded parts were replaced by *PAD*-tokens.

- *Saliency* (Simonyan et al., 2013): No special hyperparameter settings required.

- *InXGrad* (Input times gradient, Shrikumar et al., 2017): No special hyperparameter settings required.

- *Int. Grads* (Integrated Gradients, Sundararajan et al., 2017): We use a sequence of *PAD*-tokens as background and do 50 gradient evaluation steps per sample.

- *LIME* (Ribeiro et al., 2016): We do 50 function evaluations per sample. In each evaluation, we randomly select $5 - 13\%$ of tokens and replace them as well as the next three tokens with *PAD*-tokens. We train a linear classifier and use the resulting weights as rationale.

- *Shapley* (Shapley value sampling, Castro et al., 2009): We evaluate the token contributions for 25 feature permutations per sample. Removed tokens are replaced by *PAD*-tokens.

We use the implementations provided by (Kokhlikyan et al., 2020) for all methods. All methods have access to the ground truth label and therefore do not have to rely on a correct classifier prediction.

For methods that produce scores for each entry of the embedding vector, we report results for two different methods of combining these scores to single values per token, with one being taking the vector norm (results are reported for the L1 norm, but we did not see a significant difference for the L2 norm), and the other one being summing over the resulting scores (indicated by subscript n and s in Table 1, respectively). In the latter case, the resulting value was negated if the target label is 0.

To evaluate the token F1 score and the IoU F1 score, we need to create a binary mask from the continuous scores produced by the different interpretability methods. We do this by selecting the top-scoring words as rationale, with the percentage of words that are selected being decided by a Nadaraya-Watson kernel regression model using an RBF kernel. The input to the kernel regression for a given sample is the percentage of words that have a score greater than a fixed threshold (a hyperparameter, here set to 0.1), while the output is the percentage of words to be selected as rationale. As we use the rationales from (DeYoung et al., 2020) (who only annotated 200 samples) for our experiment, we do not have access to a separate training set to train the kernel regression, so we resort to a leave-one-out scheme to use the same set for training and testing.

For the faithfulness evaluation, we note that we deviate from the common practice of evaluating the area under the curve (AUC) (e.g., used by Petsiuk et al., 2018) and instead take the average over the tested range of values. We do this, to accommodate for the possibility of negative scores in the sufficiency calculation, which undermine the theoretical foundation of the AUC. We also adapt the comprehensiveness calculation accordingly for consistency.

A.2 ImageNet Explanations

We use MaRC with the optimization objective given by Equation 9. We use pretrained ResNet-101 (He et al., 2016) and vision transformer ViT-B/16 (Kolesnikov et al., 2021, input image size=384) models and use the following hyperparameter setting for MaRC:

- ResNet-101: We set $\alpha_\lambda = 0.6$, $\alpha_\sigma = 1.2$, $\alpha_{\text{NB}} = 10$ and initialize w and σ uniformly to 0.5 and 1.2 respectively. As ResNet models seem to treat uniformly colored images as uninformative, we chose B to be a set containing a black image, a white image, and an image with the mean color from the dataset. As the scoring function, we chose the log of a combination of the softmax output of c (weighted by 0.9) and the sigmoid of the logit of c (weighted by 0.1).

- Vit-B/16: We set $\alpha_\lambda = 0.25$, $\alpha_\sigma = 1.2$, $\alpha_{\text{NB}} = 10$ and initialize w and σ uniformly to 0.5 and 1.2 respectively. As the vision transformer often seems to interpret the uniformly colored backgrounds as indicative of specific classes, we instead opted to use a blurred version of the input image as b. As the scoring function, we use the log-likelihood of c.

All visualizations and experiments were, if not stated otherwise, conducted with these hyperparameter settings. We compared MaRC to the following methods:

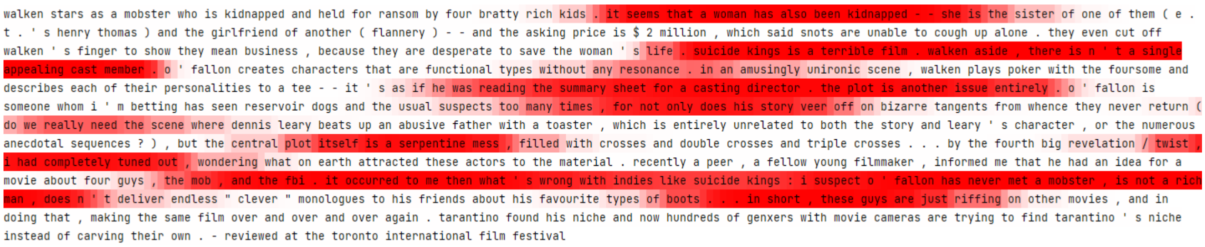
- M-Perturb (Fong and Vedaldi, 2017): We used the original implementation and parameter settings by Fong and Vedaldi (2017) with minor adaptations to work with pytorch.
- *Grad-CAM* (Selvaraju et al., 2017): We used the implementation by Gildenblat and contributors (2021).
- *Exc-BP* (Excitation Backpropagation, Zhang et al., 2016): We used the implementation available at https://github.com/greydanus/excitationbp
- *Saliency* (Simonyan et al., 2013): We used the implementation by Kokhlikyan et al. (2020).
- *InputXGrad* (Input times gradient, Shrikumar et al., 2017): We used the implementation by Kokhlikyan et al. (2020).
- *Guided BP* (Springenberg et al., 2015): We used the implementation by Kokhlikyan et al. (2020).
- *Intgr. Grads* (Integrated Gradients, Sundararajan et al., 2017): We used the implementation by Kokhlikyan et al. (2020). As background, a blurred version of the input image was used, as this produced optimal results. For each sample, 100 gradient evaluation steps were performed.
- *Occlusion* (Zeiler and Fergus, 2014): We used the implementation by Kokhlikyan et al. (2020). We occluded patches of $\frac{1}{9}$ of the input image size and used a stride of $\frac{1}{56}$ of the input image size. Occluded patches were replaced by a blurred version of the input image.
- *Attention* (Yuan et al., 2021): We used the implementation by Yuan et al. (2021).
- *Attribution* (Chefer et al., 2021b): We used the implementation by Yuan et al. (2021).
- *Rollout* (Abnar and Zuidema, 2020): We used the implementation by Yuan et al. (2021).
- *TAM* (Transition attention maps, Yuan et al., 2021): We used the implementation by Yuan et al. (2021).

All methods had access to the ground-truth label to create the explanation. The random sample of ImageNet validation images that was used in this experiment can be viewed at the GitHub page given at the beginning of Appendix A.

To evaluate model faithfulness, we used Equation 10 and 11, with the difference that the evaluation for each sample and percentage was performed with respect to four uninformative inputs to prevent skewed evaluation results for methods that only work well with respect to one background. The four backgrounds we used were a white image, a black image, an image of the mean color of the dataset, as well as a blurred version of the input image.

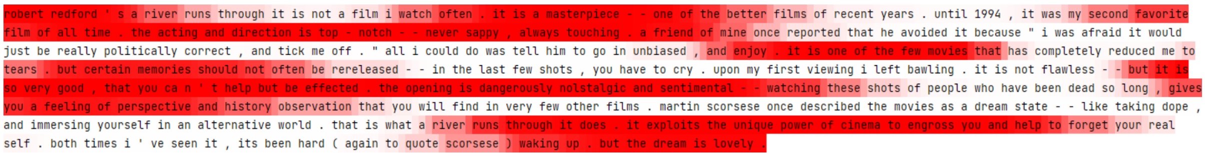
B Movie Review Rationale Examples

This section includes four additional exemplary rationales created by MaRC on movie reviews from (Zaidan et al., 2007).



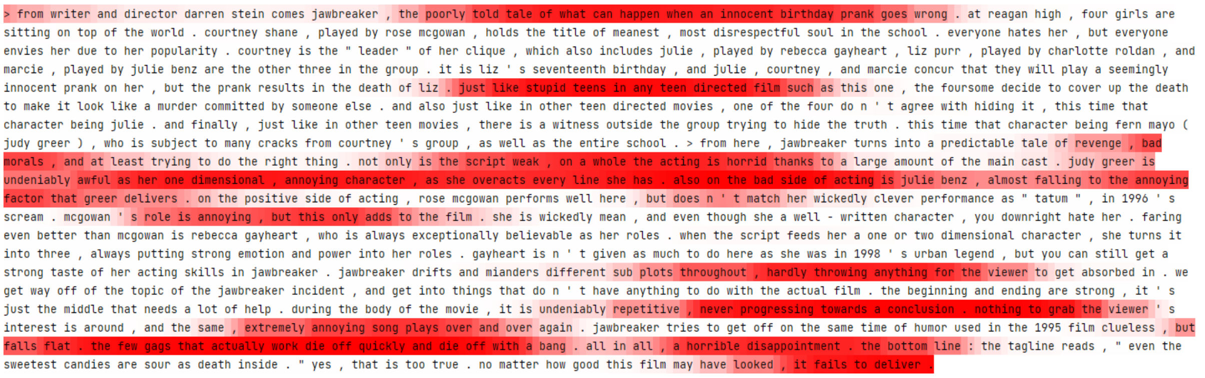
walken stars as a mobster who is kidnapped and held for ransom by four bratty rich kids . it seems that a woman has also been kidnapped - - she is the sister of one of them (e . t . ' s henry thomas) and the girlfriend of another (flannery) - - and the asking price is $ 2 million , which said snots are unable to cough up alone . they even cut off walken ' s finger to show they mean business , because they are desperate to save the woman ' s life . suicide kings is a terrible film . walken aside , there is n ' t a single appealing cast member . o ' fallon creates characters that are functional types without any resonance . in an amusingly unironic scene , walken plays poker with the foursome and describes each of their personalities to a tee - - it ' s as if he was reading the summary sheet for a casting director . the plot is another issue entirely . o ' fallon is someone whom i ' m betting has seen reservoir dogs and the usual suspects too many times , for not only does his story veer off on bizarre tangents from whence they never return (do we really need the scene where dennis leary beats up an abusive father with a toaster , which is entirely unrelated to both the story and leary ' s character , or the numerous anecdotal sequences ?) , but the central plot itself is a serpentine mess , filled with crosses and double crosses and triple crosses . . . by the fourth big revelation / twist , i had completely tuned out , wondering what on earth attracted these actors to the material . recently a peer , a fellow young filmmaker , informed me that he had an idea for a movie about four guys , the mob , and the fbi . it occurred to me then what ' s wrong with indies like suicide kings : i suspect o ' fallon has never met a mobster , is not a rich man , does n ' t deliver endless " clever " monologues to his friends about his favourite types of boots . . . in short , these guys are just riffing on other movies , and in doing that , making the same film over and over and over again . tarantino found his niche and now hundreds of genxers with movie cameras are trying to find tarantino ' s niche instead of carving their own . - reviewed at the toronto international film festival

Figure 4: Rationale for a negative movie review.



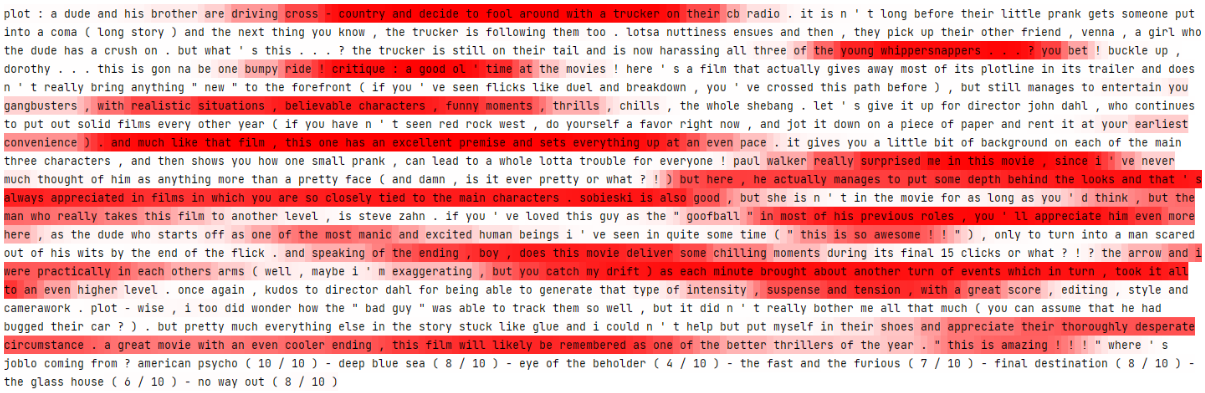
robert redford ' s a river runs through it is not a film i watch often . it is a masterpiece - - one of the better films of recent years . until 1994 , it was my second favorite film of all time . the acting and direction is top - notch - - never sappy , always touching . a friend of mine once reported that he avoided it because " i was afraid it would just be really politically correct , and tick me off . " all i could do was tell him to go in unbiased , and enjoy . it is one of the few movies that has completely reduced me to tears . but certain memories should not often be rereleased - - in the last few shots , you have to cry . upon my first viewing i left bawling . it is not flawless - - but it is so very good , that you ca n ' t help but be effected . the opening is dangerously nolstalgic and sentimental - - watching these shots of people who have been dead so long , gives you a feeling of perspective and history observation that you will find in very few other films . martin scorsese once described the movies as a dream state - - like taking dope , and immersing yourself in an alternative world . that is what a river runs through it does . it exploits the unique power of cinema to engross you and help to forget your real self . both times i ' ve seen it , its been hard (again to quote scorsese) waking up . but the dream is lovely .

Figure 5: Rationale for a positive movie review.



> from writer and director darren stein comes jawbreaker , the poorly told tale of what can happen when an innocent birthday prank goes wrong . at reagan high , four girls are sitting on top of the world . courtney shane , played by rose mcgowan , holds the title of meanest , most disrespectful soul in the school . everyone hates her , but everyone envies her due to her popularity . courtney is the " leader " of her clique , which also includes julie , played by rebecca gayheart , liz purr , played by charlotte roldan , and marcie , played by julie benz are the other three in the group . it is liz ' s seventeenth birthday , and julie , courtney , and marcie concur that they will play a seemingly innocent prank on her , but the prank results in the death of liz . just like stupid teens in any teen directed film such as this one , the foursome decide to cover up the death to make it look like a murder committed by someone else . and also just like in other teen directed movies , one of the four do n ' t agree with hiding it , this time that character being julie . and finally , just like in other teen movies , there is a witness outside the group trying to hide the truth . this time that character being fern mayo (judy greer) , who is subject to many cracks from courtney ' s group , as well as the entire school . > from here , jawbreaker turns into a predictable tale of revenge , bad morals , and at least trying to do the right thing . not only is the script weak , on a whole the acting is horrid thanks to a large amount of the main cast . judy greer is undeniably awful as her one dimensional , annoying character , as she overacts every line she has . also on the bad side of acting is julie benz , almost falling to the annoying factor that greer delivers . on the positive side of acting , rose mcgowan performs well here , but does n ' t match her wickedly clever performance as " tatum " , in 1996 ' s scream . mcgowan ' s role is annoying , but this only adds to the film . she is wickedly mean , and even though she a well - written character , you downright hate her . faring even better than mcgowan is rebecca gayheart , who is always exceptionally believable as her roles . when the script feeds her a one or two dimensional character , she turns it into three , always putting strong emotion and power into her roles . gayheart is n ' t given as much to do here as she was in 1998 ' s urban legend , but you can still get a strong taste of her acting skills in jawbreaker . jawbreaker drifts and mianders different sub plots throughout , hardly throwing anything for the viewer to get absorbed in . we get way off of the topic of the jawbreaker incident , and get into things that do n ' t have anything to do with the actual film . the beginning and ending are strong , it ' s just the middle that needs a lot of help . during the body of the movie , it is undeniably repetitive , never progressing towards a conclusion . nothing to grab the viewer ' s interest is around , and the same , extremely annoying song plays over and over again . jawbreaker tries to get off on the same time of humor used in the 1995 film clueless , but falls flat . the few gags that actually work die off quickly and die off with a bang . all in all , a horrible disappointment . the bottom line : the tagline reads , " even the sweetest candies are sour as death inside . " yes , that is too true . no matter how good this film may have looked , it fails to deliver .

Figure 6: Rationale for a negative movie review.



plot : a dude and his brother are driving cross - country and decide to fool around with a trucker on their cb radio . it is n ' t long before their little prank gets someone put into a coma (long story) and the next thing you know , the trucker is following them too . lotsa nuttiness ensues and then , they pick up their other friend , venna , a girl who the dude has a crush on . but what ' s this . . . ? the trucker is still on their tail and is now harassing all three of the young whippersnappers . . . ? you bet ! buckle up , dorothy . . . this is gon na be one bumpy ride ! critique : a good ol ' time at the movies ! here ' s a film that actually gives away most of its plotline in its trailer and does n ' t really bring anything " new " to the forefront (if you ' ve seen flicks like duel and breakdown , you ' ve crossed this path before) , but still manages to entertain you gangbusters , with realistic situations , believable characters , funny moments , thrills , chills , the whole shebang . let ' s give it up for director john dahl , who continues to put out solid films every other year (if you have n ' t seen red rock west , do yourself a favor right now , and jot it down on a piece of paper and rent it at your earliest convenience) . and much like that film , this one has an excellent premise and sets everything up at an even pace . it gives you a little bit of background on each of the main three characters , and then shows you how one small prank , can lead to a whole lotta trouble for everyone ! paul walker really surprised me in this movie , since i ' ve never much thought of him as anything more than a pretty face (and damn , is it ever pretty or what ? !) but here , he actually manages to put some depth behind the looks and that ' s always appreciated in films in which you are so closely tied to the main characters . sobieski is also good , but she is n ' t in the movie for as long as you ' d think , but the man who really takes this film to another level , is steve zahn . if you ' ve loved this guy as the " goofball " in most of his previous roles , you ' ll appreciate him even more here , as the dude who starts off as one of the most manic and excited human beings i ' ve seen in quite some time (" this is so awesome ! ! ") , only to turn into a man scared out of his wits by the end of the flick . and speaking of the ending , boy , does this movie deliver some chilling moments during its final 15 clicks or what ? ! ? the arrow and i were practically in each others arms (well , maybe i ' m exaggerating , but you catch my drift) as each minute brought about another turn of events which in turn , took it all to an even higher level . once again , kudos to director dahl for being able to generate that type of intensity , suspense and tension , with a great score , editing , style and camerawork . plot - wise , i too did wonder how the " bad guy " was able to track them so well , but it did n ' t really bother me all that much (you can assume that he had bugged their car ?) . but pretty much everything else in the story stuck like glue and i could n ' t help but put myself in their shoes and appreciate their thoroughly desperate circumstance . a great movie with an even cooler ending , this film will likely be remembered as one of the better thrillers of the year . " this is amazing ! ! ! " where ' s joblo coming from ? american psycho (10 / 10) - deep blue sea (8 / 10) - eye of the beholder (4 / 10) - the fast and the furious (7 / 10) - final destination (8 / 10) - the glass house (6 / 10) - no way out (8 / 10)

Figure 7: Rationale for a positive movie review.

C ImageNet Mask Comparison

This section includes a more extensive comparison of masks created by different interpretability methods on the selection of images used by Fong and Vedaldi (2017). For MaRC, an additional visualization is added to more easily see the unmasked regions in the image. For an overview of methods as well as hyperparameter settings, compare Appendix A.

C.1 ResNet-101

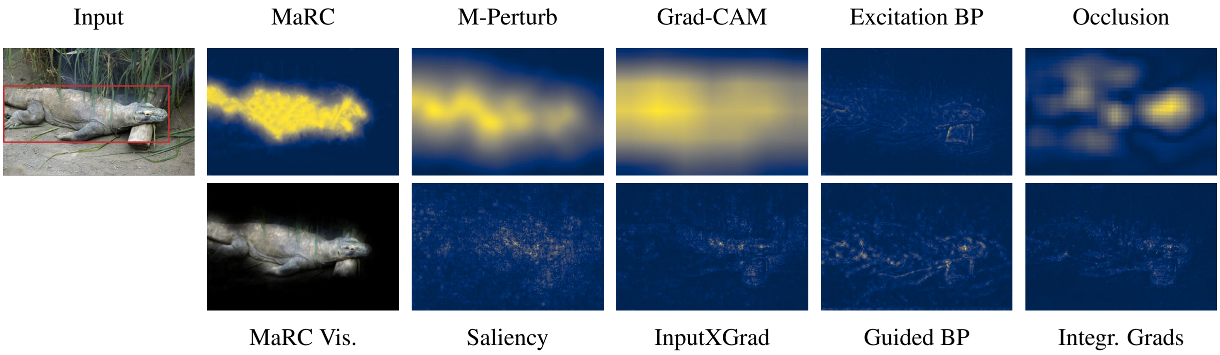


Figure 8: Komodo dragon

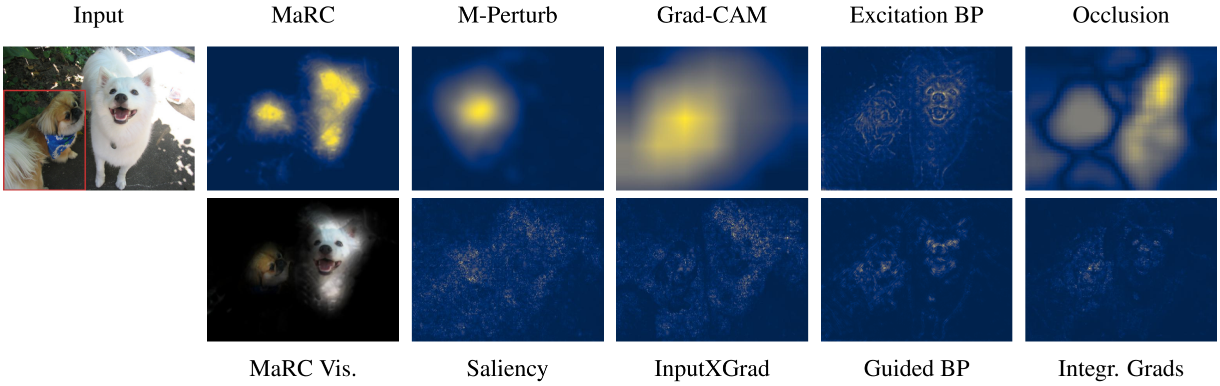


Figure 9: Pekinese

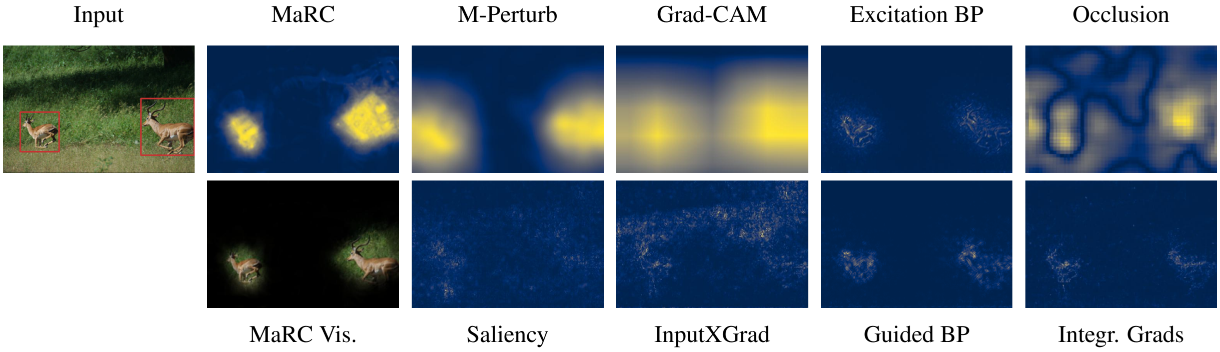


Figure 10: Impala

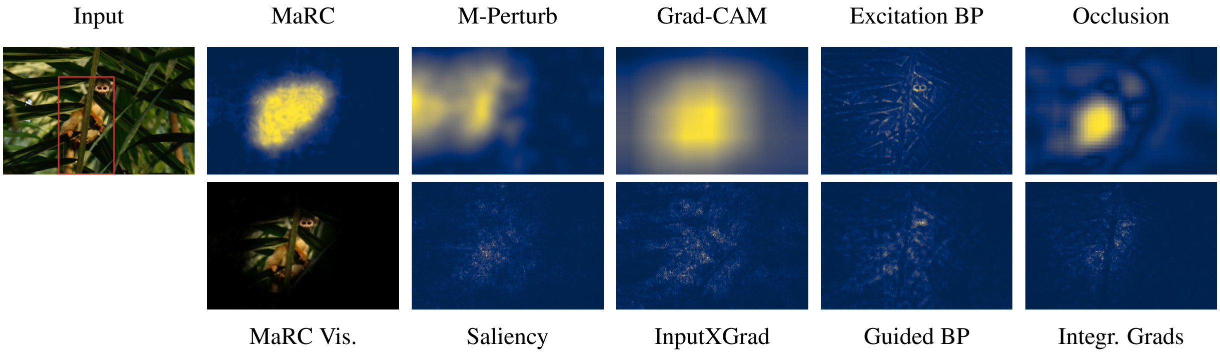


Figure 11: Squirrel monkey

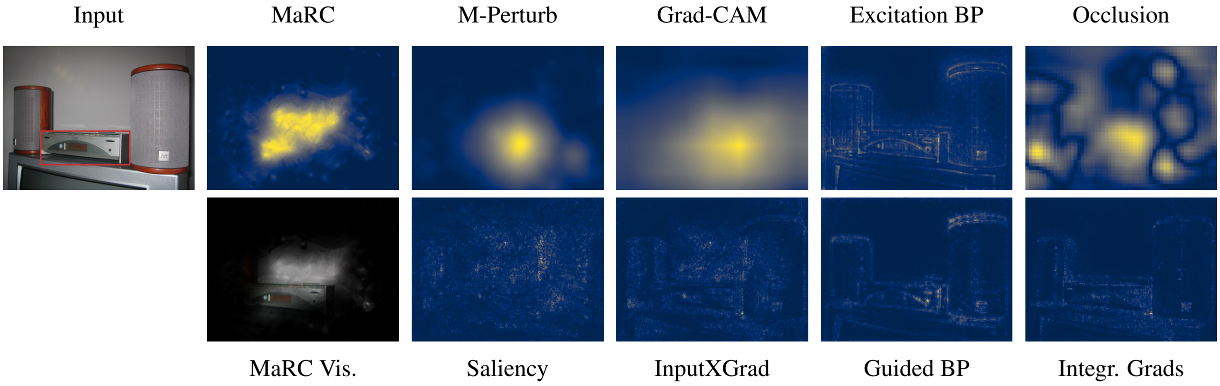


Figure 12: CD player

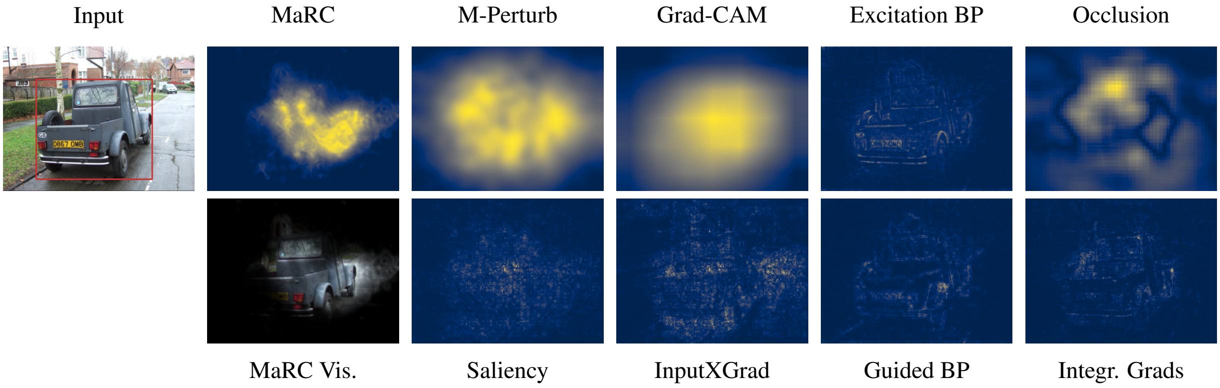


Figure 13: Pickup

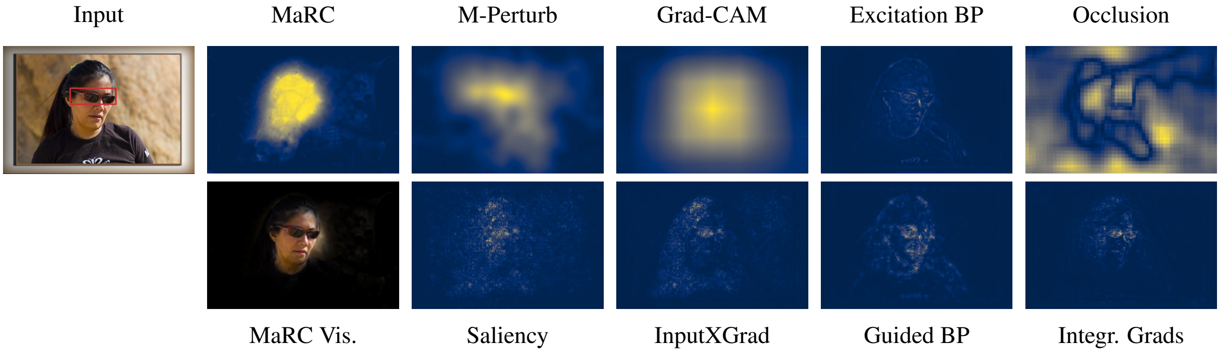


Figure 14: Sunglasses

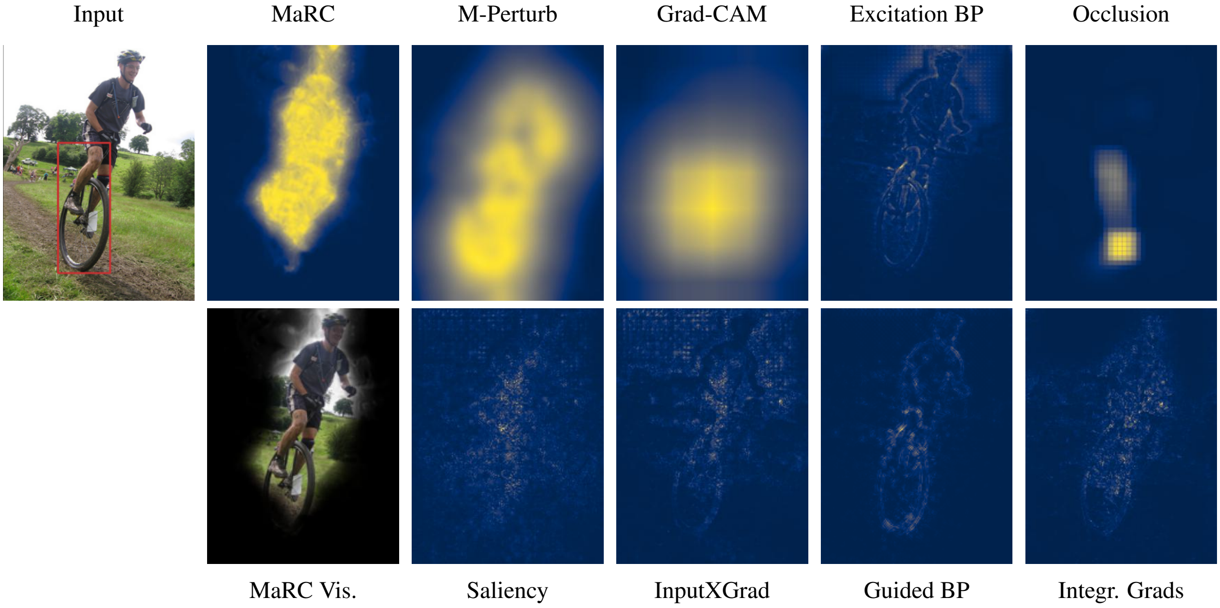


Figure 15: Unicycle

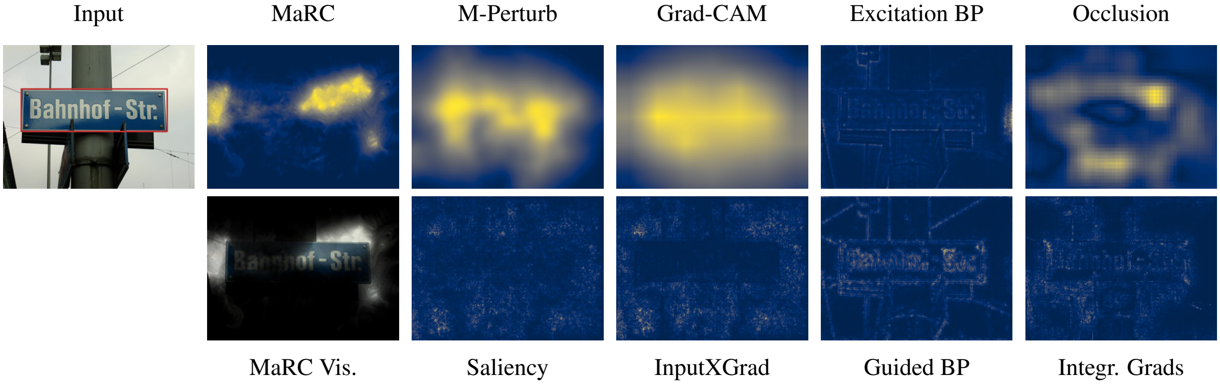


Figure 16: Street sign

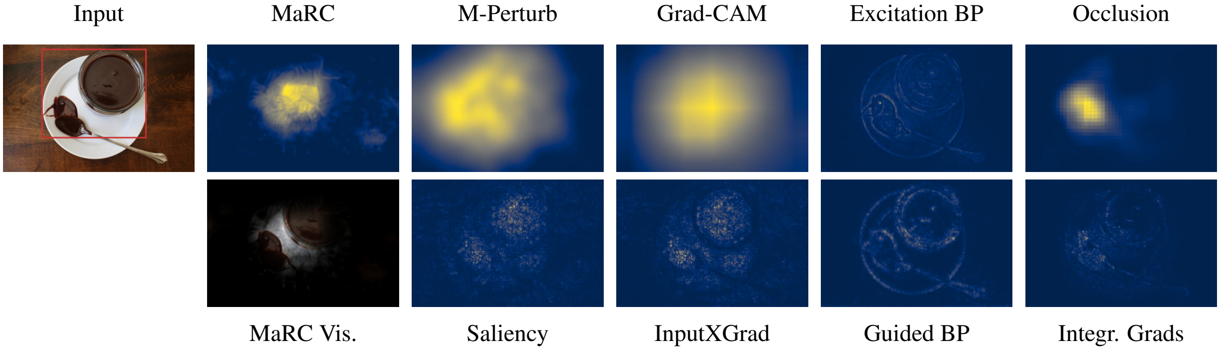


Figure 17: Chocolate sauce

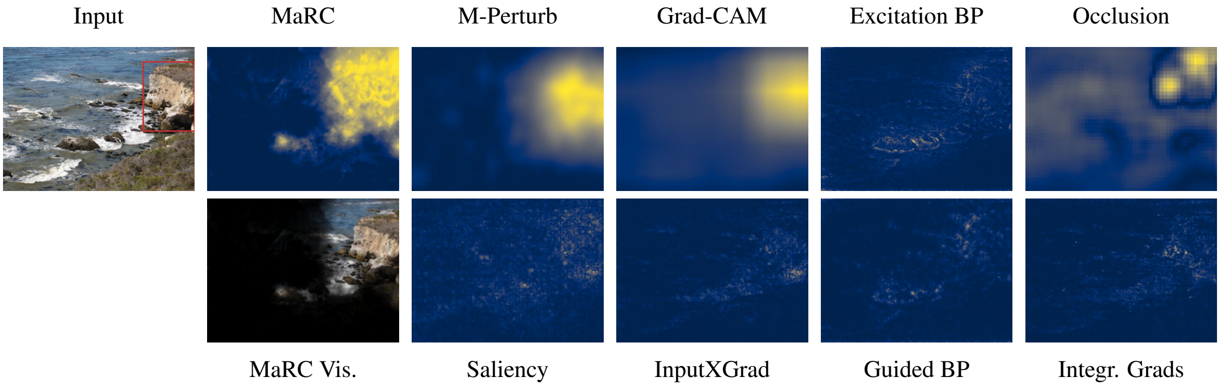


Figure 18: Cliff

C.2 ViT-B/16

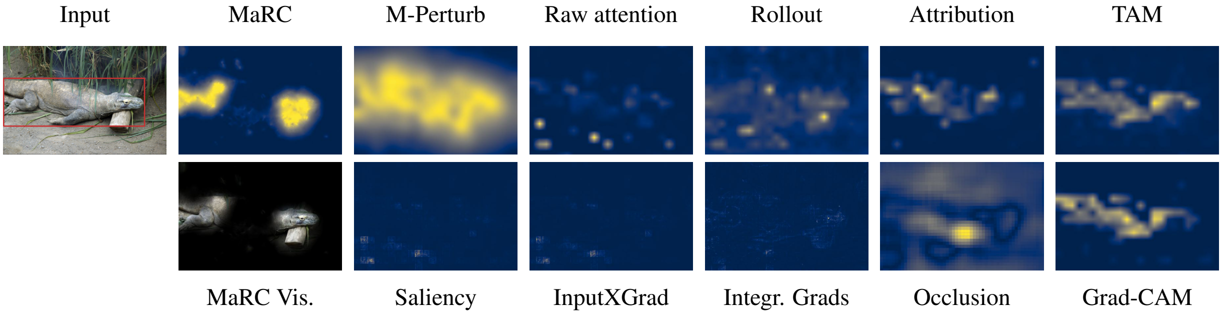


Figure 19: Komodo dragon

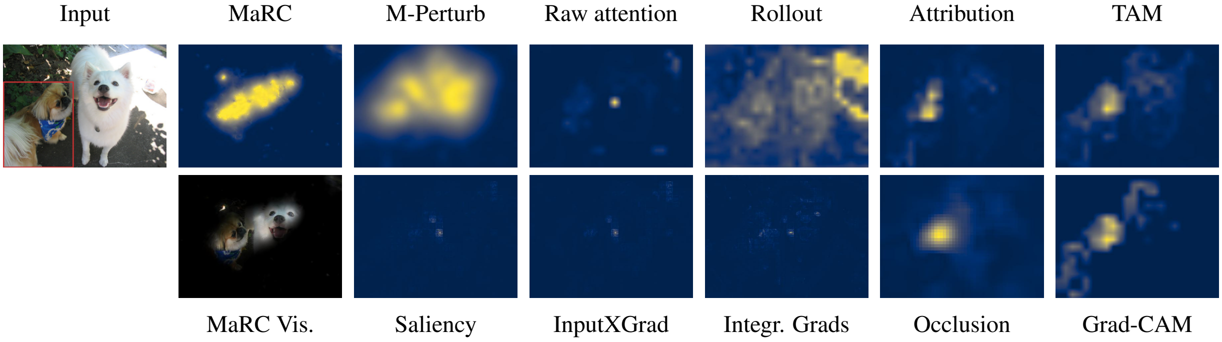


Figure 20: Pekinese

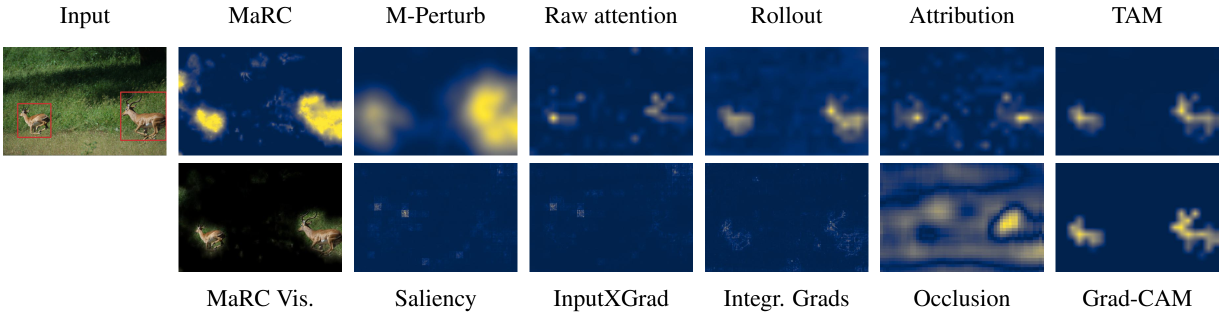


Figure 21: Impala

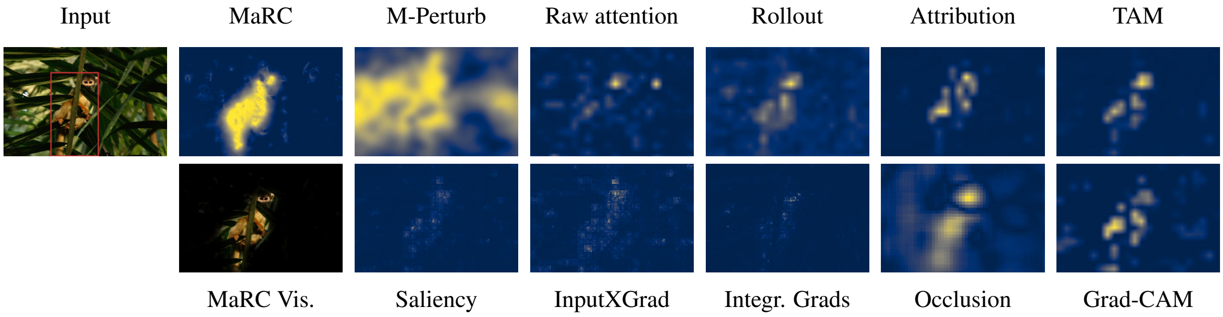


Figure 22: Squirrel monkey

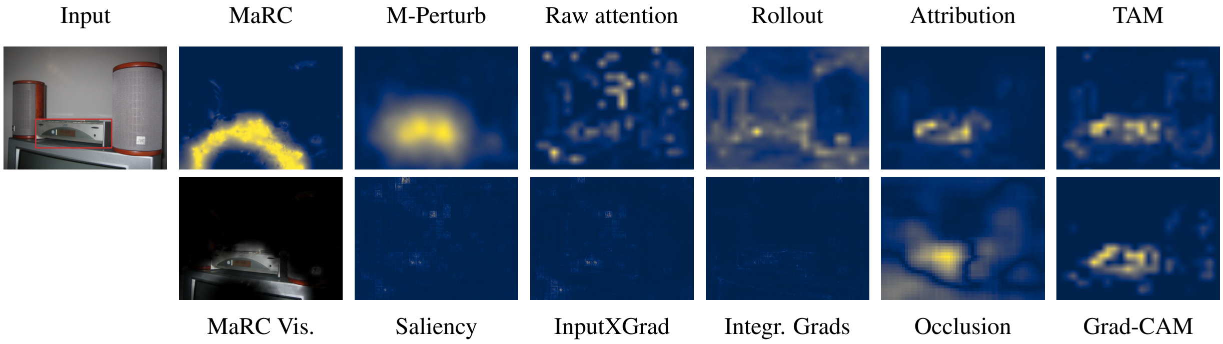


Figure 23: CD player

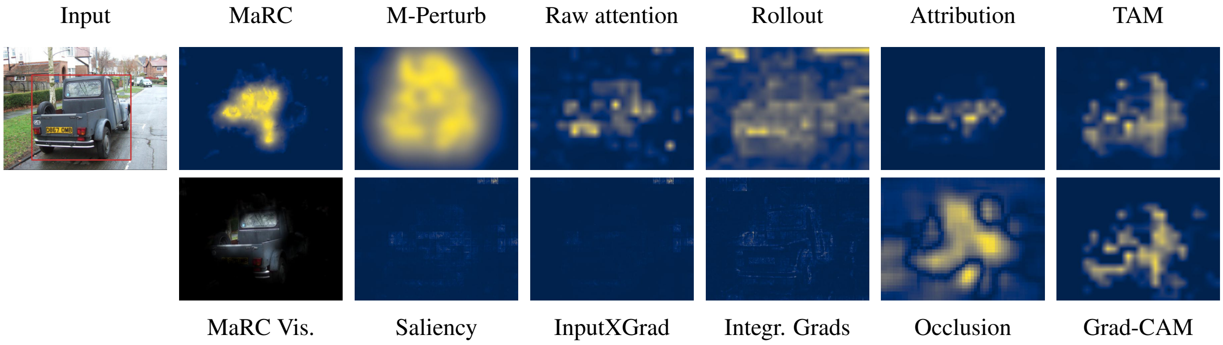


Figure 24: Pickup

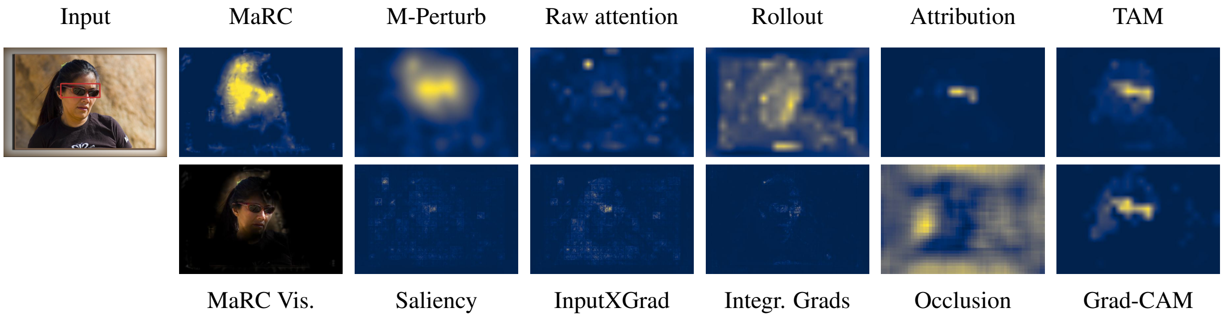


Figure 25: Sunglasses

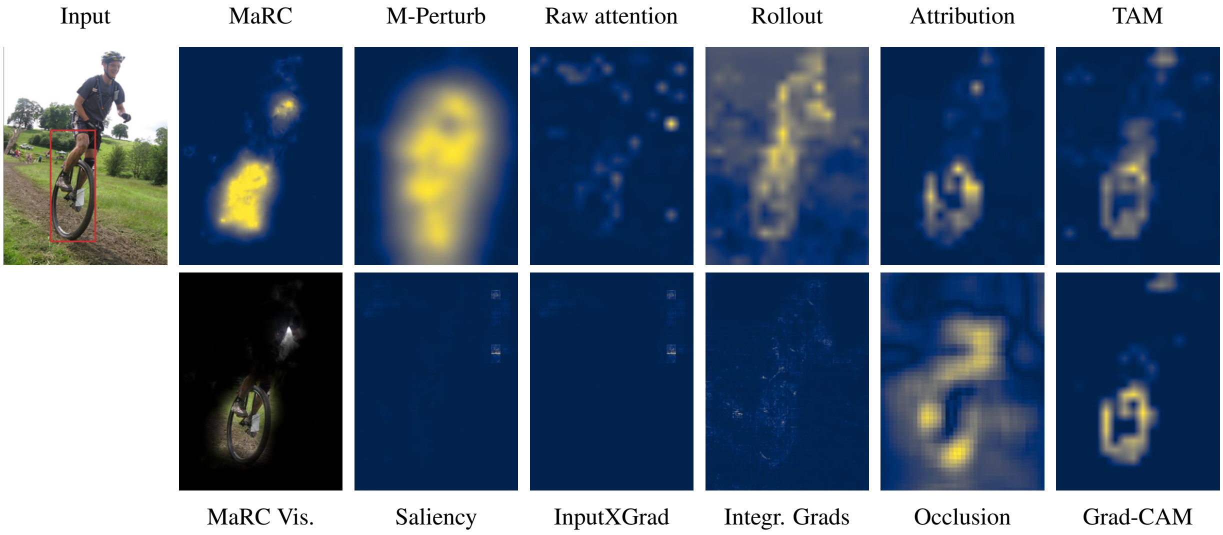


Figure 26: Unicycle

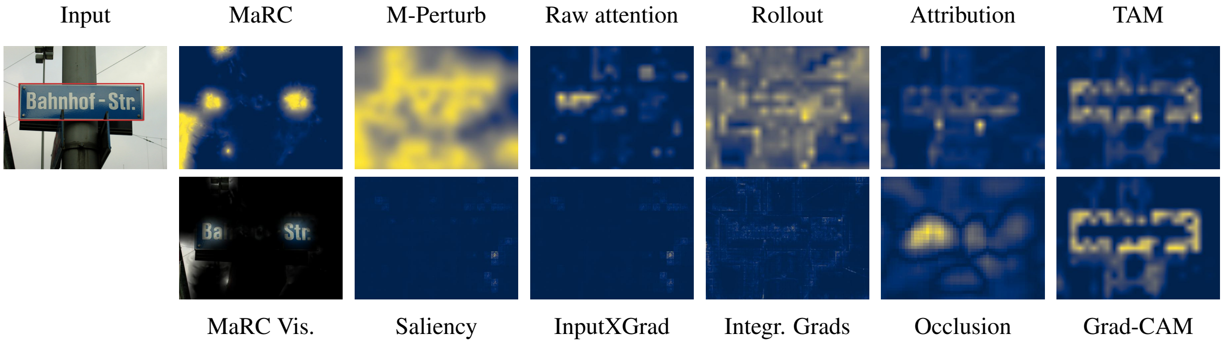


Figure 27: Unicycle

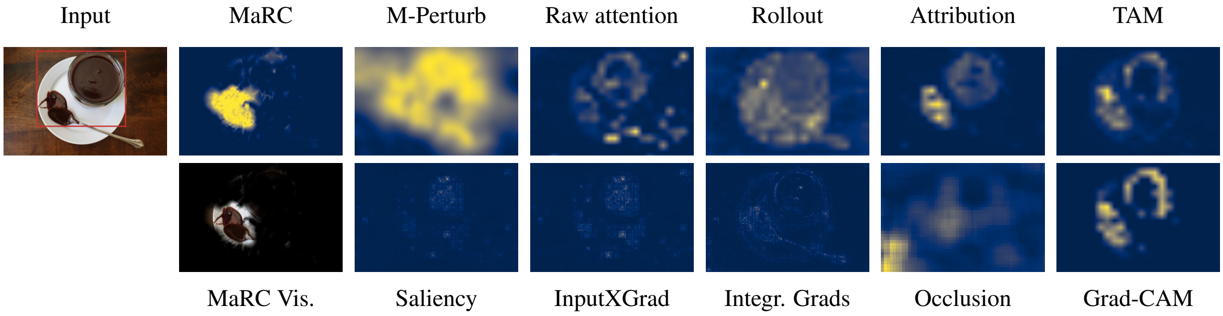


Figure 28: Chocolate sauce

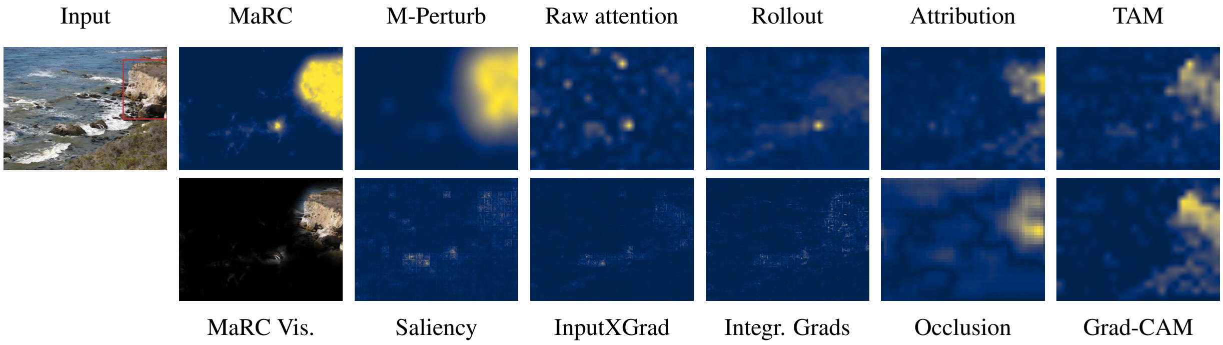


Figure 29: Cliff

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☒ A2. Did you discuss any potential risks of your work?
The method developed in our study does not create any risks.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3,4,5,B,C

☑ B1. Did you cite the creators of artifacts you used?
3,4,5,A

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The license term was not central to the paper as the used scientific artifacts were used according to it, and we do not redistribute these scientific artifacts.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We used all scientific artifacts only for scientific purposes and do not redistribute them, as well as any derivatives of them. Therefore, informing readers about the intended use these other artifacts is not necessary. We do not limit the use of the artifacts we create, and the further use of them is not central to our study.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We do not collect any data that uniquely identifies individual people or use offensive content.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
This question is not applicable to the type of scientific artifact we produced.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4, 5, A

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

C ☑ Did you run computational experiments?

4, 5, A, B, C

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Our study uses the most well-know types of models that exist in the literature. Therefore, an extensive discussion about these topics is not central to our study.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4, 5, A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4, 5, A

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We provide information about the metrics used, which are independent of specific code implementations. Further, we provide the code for our experiments and evaluations.

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

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.