Contrast-CAT: Contrasting Activations for Enhanced Interpretability in Transformer-based Text Classifiers

Sungmin Han[1] **Jeonghyun Lee**[1] **Sangkyun Lee**[*1]

[1]School of Cybersecurity, Korea University, Seoul, South Korea
{sungmin_15,nomar0107,sangkyun}@korea.ac.kr

Abstract

Transformers have profoundly influenced AI research, but explaining their decisions remains challenging – even for relatively simpler tasks such as classification – which hinders trust and safe deployment in real-world applications. Although activation-based attribution methods effectively explain transformer-based text classification models, our findings reveal that these methods can be undermined by class-irrelevant features within activations, leading to less reliable interpretations. To address this limitation, we propose Contrast-CAT, a novel activation contrast-based attribution method that refines token-level attributions by filtering out class-irrelevant features. By contrasting the activations of an input sequence with reference activations, Contrast-CAT generates clearer and more faithful attribution maps. Experimental results across various datasets and models confirm that Contrast-CAT consistently outperforms state-of-the-art methods. Notably, under the MoRF setting, it achieves average improvements of $\times 1.30$ in AOPC and $\times 2.25$ in LOdds over the most competing methods, demonstrating its effectiveness in enhancing interpretability for transformer-based text classification.

1 INTRODUCTION

Transformers [Vaswani et al., 2017] have achieved remarkable success in recent years, transcending both academic and industrial boundaries and becoming increasingly integrated into daily life. However, this widespread integration also heightens the risk of direct exposure to AI errors, underscoring the need to ensure the safety, security, and trustworthiness of AI systems through increased transparency [The White House, 2023, Dunietz et al., 2024, European Commission, 2024]. Consequently, developing methods for interpreting the decision-making processes of transformer-based models has become essential.

To address this need, numerous methods have been proposed for interpreting transformer-based models, particularly in text classification tasks where they have shown remarkable performance. These methods generate attribution maps that indicate the relative contributions of input tokens to a model's decisions. In Section 2, we categorize them into attention-based, LRP-based, and activation-based approaches. This work focuses on activation-based attribution, which leverages a model's activation information to produce attribution maps and has demonstrated state-of-the-art performance in attribution quality.

Activation-based attribution maps are typically derived by extracting activations from one or more layers of a neural network for a given input sequence. Then, the output gradient of the target class, with respect to these activations, is applied to isolate class-relevant features [Selvaraju et al., 2017]. However, we find that this procedure can still be influenced by class-irrelevant signals within the activations, thus limiting its ability to produce accurate, class-specific interpretations. For example, in Figure 1, panel (A) illustrates attribution maps generated by AttCAT [Qiang et al., 2022], one of the leading activation-based attribution methods, for the movie review 'It is very slow.', which is classified as negative. Ideally, the word 'slow' should register as highly relevant, with a positive attribution value in relation to the negative sentiment. However, AttCAT fails to detect this importance, whereas our proposed method, Contrast-CAT, correctly assigns the highest attribution to 'slow.'

In this paper, we introduce Contrast-CAT, a novel activation-based attribution method for transformer-based text classification. We find that existing methods often incorporate class-irrelevant signals, compromising attribution accuracy. By contrasting target activations with multiple reference activations, Contrast-CAT filters out these irrelevant features

[*]Corresponding author

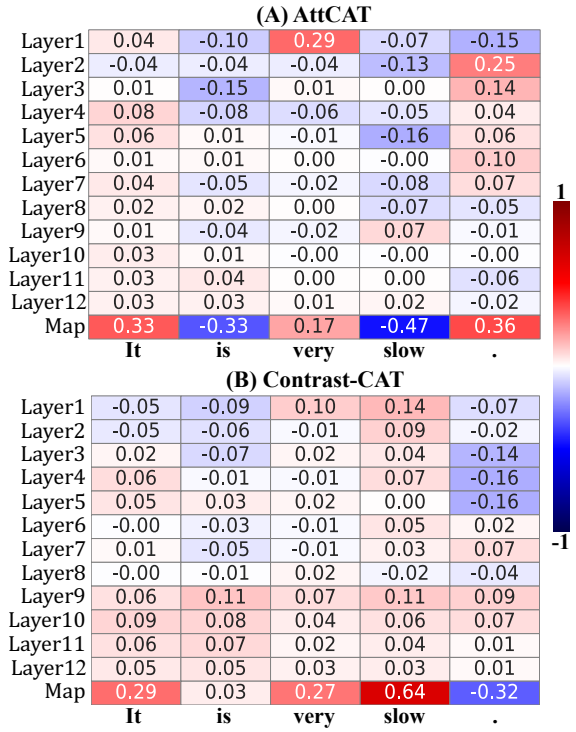

Figure 1: Heatmaps displaying attribution values from different encoder layers of the $\text{BERT}_{\text{base}}$ model for a negative review prediction. Panel A shows maps generated by AttCAT, which applies gradients directly to activations, while Panel B shows maps from Contrast-CAT, which applies gradients to activation contrast information. Values closer to 1 (red) indicate stronger contributions to the negative prediction.

and produces high-quality token-level attribution maps. Extensive experiments show that Contrast-CAT consistently outperforms state-of-the-art approaches, achieving average improvements of $\times 1.30$ and $\times 2.25$ in AOPC and LOdds under the MoRF setting, and $\times 1.34$ and $\times 1.03$ under the LeRF setting, compared to the best competitors.

2 RELATED WORK

We describe attribution methods for interpreting transformer-based text classification models, categorizing them into attention-, LRP-, and activation-based approaches.

Attention-based Attribution Attention-based attribution methods rely on attention scores, a key component of transformers. Under the assumption that input tokens with high attention scores significantly influence model outputs, numerous studies [Martins and Astudillo, 2016, Clark et al., 2019, Abnar and Zuidema, 2020, Modarressi et al., 2022, Mohebbi et al., 2023] have employed attention scores for interpretative purposes of a model. Specifically, Abnar and Zuidema [2020] proposed Rollout, which integrates attention scores across multiple layers while accounting for skip connections in transformer architectures to capture information flow. Additionally, there have been many papers [Chrysostomou and Aletras, 2021, Barkan et al., 2021] that introduce the gradient of attention weight for interpretation. Despite advances in attention-based methods, significant debate remains about whether attention scores truly reflect the relevance of model predictions, as highlighted in [Jain and Wallace, 2019, Wiegreffe and Pinter, 2019].

LRP-based Attribution Layer-wise relevance propagation (LRP) [Bach et al., 2015] is a technique for backpropagating relevance scores through a neural network, with the scores reflecting our specific interest in the model's prediction. Building on LRP, several studies have derived explanations for model behavior [Gu et al., 2018, Voita et al., 2019, Chefer et al., 2021]. In [Voita et al., 2019], LRP was partially used to determine the most important attention heads within a specific transformer's encoder layer, utilizing relevance scores for the attention weights. Chefer et al. [2021] introduces TransAtt, which propagates relevance scores through all layers of a transformer, combining these scores with gradients of the attention weights and utilizing the Rollout technique for multi-layer integration. However, LRP-based methods are limited by certain assumptions, known as the LRP rules, designed to uphold the principle of relevance conservation [Montavon et al., 2019].

Activation-based Attribution In contrast to the methods discussed above, activation-based attribution primarily relies on activation information from each layer of a transformer model. These methods are based on core ideas originally developed for convolutional neural networks (CNNs), which have been shown to be effective for generating high-quality interpretations with simple implementations and broad versatility [Selvaraju et al., 2017, Wang et al., 2020, Han et al., 2022, Lee and Han, 2022]. In [Qiang et al., 2022], the authors introduced AttCAT as the first adaptation of Grad-CAM [Selvaraju et al., 2017], one of the most popular activation-based methods for CNNs, to interpret the decisions of transformer-based text classification models. AttCAT generates token-level attribution maps by merging activations and their gradients in relation to the model's predictions, following Grad-CAM's essential approach, which uses gradients to reflect class-relevant information. Similarly, Englebert et al. [2023] introduced TIS adapting Score-CAM [Wang et al., 2020]: TIS uses the centroids of activation clusters identified from the activation from all layers to compute relevance scores in a manner akin to Score-CAM.

Although there are attribution methods for transformer-based text classification models that use gradients to extract class-relevant features from activations, no approach has yet focused on filtering out class-irrelevant features through activation contrasting to improve token-level attribution quality.

3 PRELIMINARY

Problem Statement Consider a pre-trained transformer-based model as a function f processing input tokens $x := \{x_i\}_{i=1}^T$, where T is the length of the input sequence, and each token is denoted as $x_i \in \mathbb{R}^n$. Our objective is to generate a token-level attribution map $I(x) := \{I(x)_i\}_{i=1}^T$, where $I(x)_i$ represents the relevance score of each input token x_i regarding the output $f(x)$.

Transformers Let us consider a transformer-based model which is composed of L stacked layers of identical structure. We denote that the ℓ-th layer outputs an activation sequence $A^\ell := \{A_i^\ell\}_{i=1}^T$ that corresponds to input tokens, where $A_i^\ell \in \mathbb{R}^n$. Each layer computes its output by combining the output from the attention layer with the previous layer's activation, where the attention layer calculates the attention scores:

$$\alpha^{\ell,h} := \text{softmax}\left(Q^{\ell,h}(A^{\ell-1}) \cdot K^{\ell,h}(A^{\ell-1})^T/\sqrt{d}\right). \quad (1)$$

Here, $Q^{\ell,h}(\cdot)$, $K^{\ell,h}(\cdot)$, and $V^{\ell,h}(\cdot)$ are the transformations for computing the query, key, and value of the ℓ-th layer's h-th head, respectively, and d is a scaling factor. $\alpha^{\ell,h} \in \mathbb{R}^{T\times T}$ refers to the attention map of the h-th head, which contains attention scores, where $h = 1 \dots H$. We denote by $\tilde{A}^{\ell,h}$ the output of the h-th attention head in the ℓ-th layer:

$$\tilde{A}^{\ell,h} := \alpha^{\ell,h} \cdot V^{\ell,h}(A^{\ell-1}).$$

The outputs from multiple attention heads are concatenated and then combined using a fully connected layer with the skip connection: $\hat{A}^\ell := \text{Concat}(\tilde{A}^{\ell,1}, \tilde{A}^{\ell,2}, \dots, \tilde{A}^{\ell,H}) \cdot \tilde{W}^\ell + A^{\ell-1}$, where $\tilde{W}^\ell$ is the weight of the fully connected layer. Finally, the ℓ-th layer's output $A^\ell \in \mathbb{R}^{T\times n}$ is computed using a feed-forward layer and skip connection:

$$A^\ell = \hat{A}^\ell \cdot W^\ell + \hat{A}^\ell, \quad (2)$$

where $W^\ell \in \mathbb{R}^{n\times n}$ is the weight for the feed-forward layer. We have omitted bias parameters and layer normalization in the above expressions for simplicity.

4 METHODOLOGY

We introduce Contrast-CAT, a *token-level*, *activation-based* attribution framework tailored to *transformer* models.

4.1 ATTRIBUTION MAP

Let $x := \{x_i\}_{i=1}^T$ be a sequence of T tokens, and let $f_c(x)$ denote the model's score for the target class c. For each token x_i ($i = 1, \dots, T$), Contrast-CAT defines its *attribution* with respect to a *contrastive reference* R as:

$$I_R(x)_i := \sum_{\ell=1}^{L} \hat{\alpha}_i^\ell \sum_{j=1}^{n} \left(\frac{\partial f_c(x)}{\partial A_i^\ell} \odot \left(A_i^\ell - R_i^\ell\right)\right)_j. \quad (3)$$

Here,

- $A_i^\ell \in \mathbb{R}^n$ is the activation for token x_i at layer ℓ,
- $\frac{\partial f_c(x)}{\partial A_i^\ell} \in \mathbb{R}^n$ is the gradient of $f_c(x)$ w.r.t. A_i^ℓ,
- R_i^ℓ is a *reference activation* for token i chosen from a reference token sequence r such that $f_c(r) < \gamma$,
- $\odot$ denotes element-wise multiplication,
- $\hat{\alpha}_i^\ell$ is the *averaged attention* of token i at layer ℓ.

In essence, $\left(A_i^\ell - R_i^\ell\right)$ *contrasts* the target activation against one that does not strongly activate class c, thereby removing non-target signals (class-irrelevant features). The factor $\frac{\partial f_c(x)}{\partial A_i^\ell}$ highlights the parts of the activation that actually affect the model's output, while $\hat{\alpha}_i^\ell$ weights these elements by how much the transformer attends to token i.

Figure 2 provides a simplified illustration of the attribution map construction process for Contrast-CAT.

4.2 COMPONENT DETAILS AND MOTIVATION

Token-Level Activations A_i^ℓ. Transformers represent each token x_i as a vector in each layer ℓ. By working at the *token level*, Contrast-CAT directly captures the discrete, context-dependent nature of language—differentiating it from CNN-based attribution methods initially designed for spatial feature maps.

Gradients $\frac{\partial f_c(x)}{\partial A_i^\ell}$. Inspired by gradient-based interpretations, we leverage the partial derivative of $f_c(x)$ w.r.t. A_i^ℓ. This follows general insights from activation-based methods, (e.g., [Selvaraju et al., 2017]), ensuring that only components of A_i^ℓ that genuinely influence $f_c(x)$ are emphasized.

Activation Contrasting $A_i^\ell - R_i^\ell$. A key novelty of Contrast-CAT is its *contrast* operation, which computes the difference between a target activation A_i^ℓ and a *low-activation* reference R_i^ℓ. The reference R_i^ℓ is chosen from a sequence r such that $f_c(r) < \gamma$, where γ is a pre-defined small positive number ($\gamma > 0$). This choice ensures that the reference activation has a minimal response to the target class c (we set $\gamma = 10^{-3}$ in our experiments). While the use of reference or baseline activations is broadly motivated by prior works (e.g., [Lee and Han, 2022]), Contrast-CAT is the first to extend this idea to transformer-based text classification networks, applying it *across multiple transformer layers*, at the *token level*, explicitly targeting textual data. This operation highlights class-specific features that distinguish x from a weakly activating example.

Attention Weights $\hat{\alpha}_i^\ell$. Transformers distribute relevance across tokens via multi-head attention. We aggregate these attention scores into $\hat{\alpha}_i^\ell$, giving higher importance to tokens that the model itself regards as salient. Unlike purely

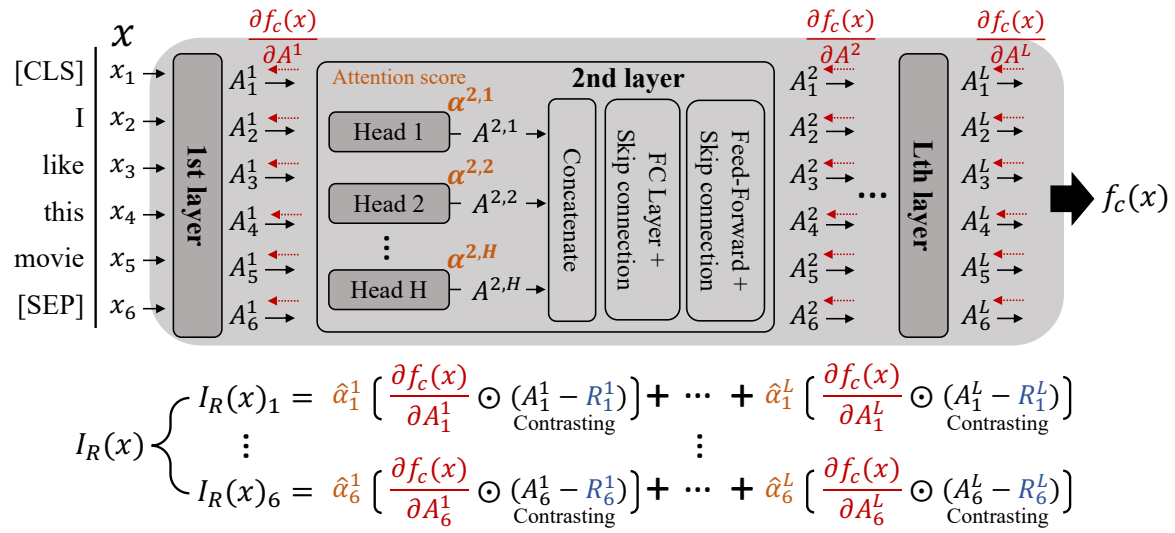

Figure 2: Construction of Contrast-CAT's attribution map. For an input token sequence x, Contrast-CAT computes an attribution map $I_R(x)$ by contrasting the *target activation* A (black) with a *reference activation* R (blue), then weighting by gradients (red) and attention (yellow).

attention-based methods (e.g., [Abnar and Zuidema, 2020]), Contrast-CAT integrates attention and gradient-based cues, offering a more robust attribution signal.

Multi-Layer Attribution Building on prior findings that transformers encode varying levels of semantic information across their layers—ranging from phrase-level details to deeper semantic features [Jawahar et al., 2019, Turton et al., 2021, Pascual et al., 2021]—we diverge from traditional activation-based attribution methods which typically rely on a single layer (e.g., [Barkan et al., 2021]). Instead, we incorporate *multi-layer* activations A^ℓ from all layers $\ell = 1, \ldots, L$ in Eq. (2), together with their layer-wise attention scores $\alpha^{\ell,h}$ in Eq. (1). This design captures *layer-specific* token semantics, and by weighting them with $\hat{\alpha}_i^\ell$, it effectively highlights the tokens most influential to the model's output across all layers.

4.3 ATTRIBUTION WITH MULTIPLE CONTRAST

Relying on a *single* reference from one class can be insufficient if the target activations $A^\ell := \{A_i^\ell\}_{i=1}^T$ encode features shared across *multiple* non-target classes. Moreover, any features that consistently remain after contrasting A^ℓ with several reference activations are more likely to represent class-specific properties. To address this, we generate a collection of attribution maps

$$D := \left\{ I_{R(r)}(x) \;\middle|\; r \in \text{training set},\ f_c(r) < \gamma \right\},$$

by repeating the procedure in Section 4.1 with *multiple* reference sequences. We cache these reference activations—one might call it a *reference library*—for use during inference. In practice, we employ 30 pre-computed references per class.

Refinement via Deletion Test Although this multi-reference approach reduces the risk of overlooking crucial class-relevant features, not all resulting maps $I_{R(r)}(x)$ are guaranteed to be reliable. We therefore *refine* Contrast-CAT by examining each map's *attribution quality* using a token-wise deletion test (e.g., [Petsiuk, 2018, Wang et al., 2020]). Specifically, we remove the top-attributed tokens one by one and record how much the model's predictive probability for class c decreases. The *average probability drop score* captures, on a token-by-token basis, how effectively a map localizes truly important tokens.

Any map with a drop score below a specified threshold ρ (set in our experiments to the mean plus one standard deviation of all drop scores) is discarded. Finally, we generate the Contrast-CAT attribution by averaging all remaining high-quality maps:

$$I(x) := \frac{1}{|M|} \sum_{I_R(x) \in M} I_R(x),$$
$$\text{where} \quad M := \left\{ I_R(x) \in D : S\big(I_R(x)\big) \geq \rho \right\}.$$

This final aggregation fuses the most credible contrastive perspectives into a single, robust token-level attribution.

5 EXPERIMENTS

Experiment Settings We implemented our method, Contrast-CAT, using PyTorch (the code is available at `https://github.com/ku-air/Contrast-CAT`). We used the $\text{BERT}_{\text{base}}$ model [Devlin et al., 2019], consisting of 12 encoder layers with 12 attention heads, as the transformer-based model for our experiments (see the supplementary material for results using other

transformer-based models). We evaluated our method on four popular datasets for text classification tasks: Amazon Polarity [Zhang et al., 2015], Yelp Polarity [Zhang et al., 2015], SST2 [Socher et al., 2013], and IMDB [Maas et al., 2011]. We reported our results using 2000 random samples from the test sets of each dataset, except for SST2, for which the entire set was used since the entire dataset had fewer than 2000 samples.

We compared our method to various attribution methods, categorized by attention-based: RawAtt, Rollout [Abnar and Zuidema, 2020], Att-grads, Att$\times$Att-grads, and Grad-SAM [Barkan et al., 2021]; LRP-based: Full LRP [Ding et al., 2017], Partial LRP [Voita et al., 2019], and TransAtt [Chefer et al., 2021]; and activation-based methods: CAT, AttCAT [Qiang et al., 2022], and TIS [Englebert et al., 2023].

Evaluation Metrics We used the area over the perturbation curve (denoted by AOPC) [Nguyen, 2018, Chen et al., 2020] and the log-odds (LOdds) [Shrikumar et al., 2017, Chen et al., 2020] metrics for assessing the faithfulness of attribution following the previous research [Qiang et al., 2022]. Faithfulness refers to the accuracy with which an attribution map's scores reflect the actual influence of each input token on the model's prediction. The AOPC and LOdds metrics are defined as follows: (1) AOPC(k) := $\frac{1}{N}\sum_{i=1}^{N}(y_i^c - \tilde{y}_i^c)$, and (2) LOdds($k$) := $\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\tilde{y}_i^c}{y_i^c}\right)$. Here, N is the total number of data points used for evaluation, and y_i^c denotes the model's prediction probability for the class c of a given input token sequence x, while $\tilde{y}_i^c$ indicates the probability after removing the top-$k\%$ of input tokens based on relevance scores from an attribution map.

To evaluate attribution quality more precisely using the AOPC and LOdds metrics while addressing inconsistencies from token removal order (i.e., removing the most relevant tokens first versus the least relevant tokens first) [Rong et al., 2022], we conducted experiments under two settings: one where tokens were removed in descending order of relevance scores (MoRF: Most Relevant First), and another in ascending order (LeRF: Least Relevant First). Consistently achieving high-quality attribution under both conditions indicates superior attribution quality. Specifically, under the MoRF setting, higher AOPC and lower LOdds indicate better attribution, while under the LeRF setting, lower AOPC and higher LOdds suggest better performance.

5.1 FAITHFULNESS OF ATTRIBUTION

Figure 3 illustrates the AOPC and LOdds values for attribution maps generated by each competing method, evaluated at various top-$k\%$ thresholds where k is increased by 10 within the range of $[10, 90]$. Table 1 provides the corresponding AUC values. Note that Figure 3 presents results for the MoRF setting only, while Table 1 includes results for both MoRF and LeRF settings. Through this evaluation, we can analyze the overall characteristics of an attribution map in terms of relevance scores of different threshold levels.

The trends in Figure 3 reveal that our method, Contrast-CAT, consistently maintains faithful attribution quality across all threshold levels and datasets compared to other methods. Table 1 further supports this, showing that Contrast-CAT achieves top-1 attribution quality under both MoRF and LeRF settings. Specifically, compared to the second-best cases, Contrast-CAT shows average improvements in AUC values of AOPC and LOdds under the MoRF setting by $\times 1.30$ and $\times 2.25$, respectively. For the LeRF setting, Contrast-CAT shows average improvements in AUC values of AOPC and LOdds by $\times 1.34$ and $\times 1.03$, respectively.

5.2 QUALITATIVE EVALUATION

Figure 4 illustrates the attribution maps generated by Contrast-CAT, TIS, and AttCAT, the top-3 ranked methods in our faithfulness evaluation, conducted under the MoRF setting (Table 1, (A) MoRF). The examples provided are from the SST2 dataset. For ease of interpretation, only tokens with relevance scores exceeding 0.5 are highlighted. As shown in the left side of Figure 4, Contrast-CAT identifies relevant tokens related to the predicted class, such as 'fails' or 'disappointment' for the negative prediction cases. For a positive prediction, in the input phrase 'rare birds have more than enough charm to make it memorable.', Contrast-CAT highlights 'enough' and 'charm' as the most relevant tokens, with 'than', 'make', 'more', and 'memorable' following in relevance. In contrast, AttCAT focuses only on 'enough' and 'memorable', missing 'charm' and 'more', while TIS identifies 'to' as the most relevant token.

5.3 THE EFFECT OF ACTIVATION CONTRAST

To evaluate the effect of our Contrast-CAT's activation contrasting, we compared the attribution quality of different versions of Contrast-CAT: the 'Random' version uses randomly selected references from individual training datasets instead of what had been outlined in Section 4.3, and the 'Same' version uses references of the same class as the target instead of different classes. The 'Same' version contrasts with our method, which leverages activations from different classes as contrastive references.

Table 2 presents AUC values of each version of Contrast-CAT, where the suggested Contrast-CAT is denoted by 'Contrasting'. The attribution quality is the worst with 'Same' and the best with 'Contrasting', which indicates that the proposed activation contrasting effectively reduces non-target signals in the activations, thereby helping to generate high-quality attribution maps.

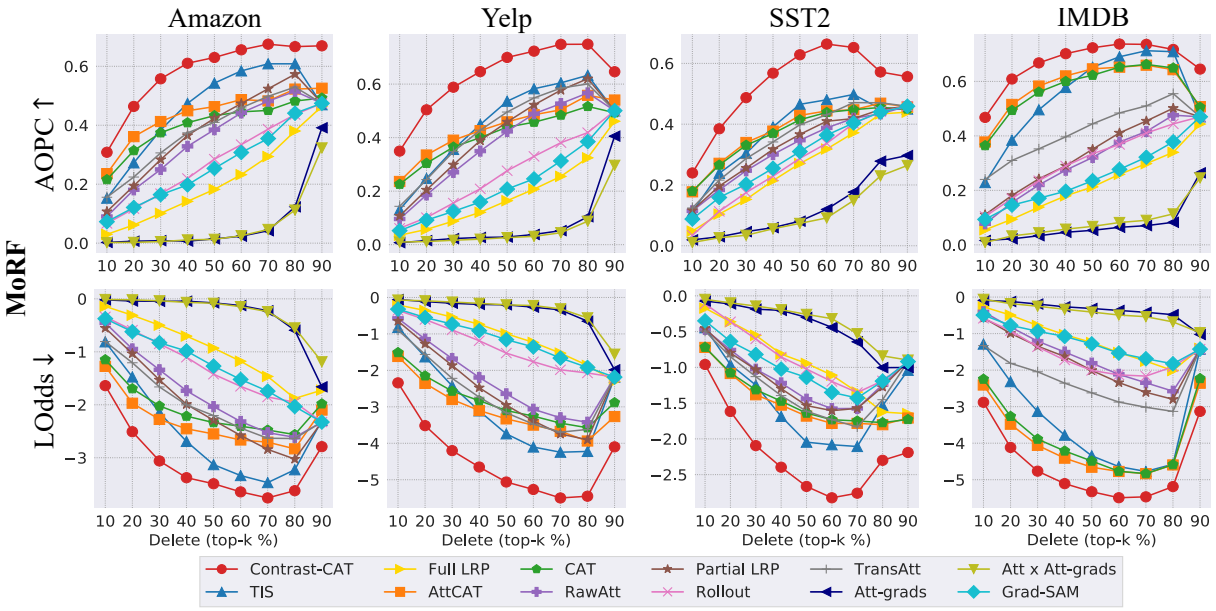

Figure 3: Quantitative comparison of the faithfulness evaluation of Contrast-CAT and other attribution methods, measured under the MoRF (Most Relevant First) setting. The arrows mean that ↑: higher is better, and ↓: lower is better.

	Class : Negative	**Class : Positive**
Input	the movie fails to live up to the sum of its parts.	rare birds has more than enough charm to make it memorable.
Contrast-CAT	the movie fails to live up to the sum of its parts .	rare birds has more than enough charm to make it memorable .
AttCAT	the movie fails to live up to the sum of its parts .	rare birds has more than enough charm to make it memorable .
TIS	the movie fails to live up to the sum of its parts .	rare birds has more than enough charm to make it memorable .
Input	my reaction in a word : disappointment.	a warm, funny, engaging film.
Contrast-CAT	my reaction in a word : disappointment .	a warm , funny , engaging , film .
AttCAT	my reaction in a word : disappointment .	a warm , funny , engaging , film .
TIS	my reaction in a word : disappointment .	a warm , funny , engaging , film .

Figure 4: Qualitative comparison of attribution quality. Relevance scores are shown with color shades: red for the highest importance, followed by orange.

5.4 CONFIDENCE OF ATTRIBUTION

If an attribution method consistently generates similar attribution maps regardless of the model's prediction, its confidence is questionable. Therefore, we conducted the confidence evaluation of the attribution methods employing the Kendall-τ rank correlation [Kendall, 1955], which is a statistical measure used to assess the similarity between two data by comparing the ranking order of their respective values. We compute an averaged rank correlation:

$$\frac{1}{N}\sum_{i=1}^{N} \text{Kendall-}\tau(P_i^c, P_i^{\hat{c}}),$$

where P_i^c is an array of token indices in descending order of relevance scores for class c in an attribution map, $P_i^{\hat{c}}$ is a similar array but for the class $\hat{c} \neq c$, and N is the total number of data points used for testing. For the choice of $\hat{c}$, we followed the settings of AttCAT as detailed in their open-source implementation, where the class immediately following the class c was chosen.

If an attribution method assigns relevance scores to tokens in distinct orders for different class predictions of the inspected model, the rank correlation is expected to be low. Table 3 presents the average rank correlation for various attribution methods tested across datasets. Cases with average rank correlation values under 0.05 are marked as '< 0.05' and highlighted: these are the cases where the attribution methods seem to work soundly – our Contrast-CAT seems to pass the test, along with Att-grads, Att$\times$Att-grads, CAT, AttCAT and TIS. In contrast, methods such as RawAtt, Rollout, and Partial LRP showed values near 1.0 consistently over the datasets, suggesting that these methods have issues generating distinct attribution over different class outcomes.

5.5 THE EFFECT OF USING MULTIPLE LAYERS

Panel (A) of Figure 5 demonstrates the effect of using multiple layers to improve the attribution quality of Contrast-CAT. The figure shows the average AUC values of AOPC and LOdds across datasets, measured under the MoRF setting.

Table 1: AUC values from the faithfulness evaluation, with (A) showing results under the MoRF (Most Relevant First) setting and (B) showing results under the LeRF (Least Relevant First) setting. The best and second-best results are highlighted in bold and underlined, respectively. The arrows mean that ↑: higher is better, and ↓: lower is better.

(A) MoRF (Most Relevant First)								
Dataset	Amazon		Yelp		SST2		IMDB	
Method	AOPC↑	LOdds↓	AOPC↑	LOdds↓	AOPC↑	LOdds↓	AOPC↑	LOdds↓
RawAtt	0.424	0.405	0.412	0.462	0.386	0.471	0.335	0.564
Rollout	0.327	0.516	0.282	0.601	0.329	0.558	0.339	0.566
Att-grads	0.061	0.749	0.059	0.754	0.132	0.691	0.061	0.759
Att×Att-grads	0.054	0.756	0.045	0.763	0.109	0.711	0.075	0.746
Grad-SAM	0.312	0.526	0.235	0.633	0.356	0.518	0.266	0.623
Full LRP	0.242	0.592	0.190	0.652	0.310	0.538	0.233	0.631
Partial LRP	0.463	0.356	0.447	0.422	0.400	0.461	0.364	0.538
TransAtt	0.461	0.366	0.473	0.404	0.432	0.428	0.458	0.455
CAT	0.482	0.341	0.440	0.383	0.452	0.382	0.632	0.215
AttCAT	0.527	0.292	0.470	<u>0.346</u>	0.461	0.372	<u>0.644</u>	<u>0.198</u>
TIS	<u>0.560</u>	<u>0.241</u>	<u>0.494</u>	0.349	<u>0.463</u>	<u>0.367</u>	0.618	0.277
Contrast-CAT	**0.703**	**0.117**	**0.687**	**0.131**	**0.654**	**0.157**	**0.738**	**0.101**
(B) LeRF (Least Relevant First)								
Dataset	Amazon		Yelp		SST2		IMDB	
Method	AOPC↓	LOdds↑	AOPC↓	LOdds↑	AOPC↓	LOdds↑	AOPC↓	LOdds↑
RawAtt	0.133	0.694	0.093	0.723	0.249	0.577	0.158	0.688
Rollout	0.166	0.670	0.130	0.687	0.373	0.448	0.126	0.711
Att-grads	0.636	0.186	0.560	0.252	0.601	0.223	0.588	0.271
Att×Att-grads	0.707	0.111	0.660	0.145	0.681	0.126	0.709	0.127
Grad-SAM	0.139	0.677	0.107	0.713	0.285	0.547	0.118	0.715
Full LRP	0.254	0.588	0.187	0.649	0.377	0.454	0.199	0.656
Partial LRP	0.122	0.700	0.088	0.725	0.237	0.585	0.134	0.701
TransAtt	0.089	0.731	<u>0.063</u>	<u>0.751</u>	0.215	0.605	<u>0.061</u>	<u>0.761</u>
CAT	0.108	0.712	0.087	0.727	0.213	0.611	0.128	0.697
AttCAT	<u>0.078</u>	<u>0.740</u>	<u>0.063</u>	0.747	<u>0.205</u>	<u>0.623</u>	0.119	0.703
TIS	0.104	0.719	0.082	0.737	0.252	0.562	0.135	0.691
Contrast-CAT	**0.058**	**0.757**	**0.048**	**0.759**	**0.147**	**0.669**	**0.047**	**0.775**

The results in panel (A) of Figure 5 indicate that the attribution quality improves as the number of layers increases, with the best performance achieved when all layers are used. Specifically, there is a $\times 1.52$ improvement in AOPC and $\times 3.05$ improvement in LOdds when using all layers compared to using only the penultimate layer. The AOPC and LOdds values tend to saturate when we use three or more layers but continue to increase as the number increases.

5.6 THE EFFECT OF MULTIPLE CONTRASTS

Panel (B) of Figure 5 illustrates the impact of increasing the number of references for multiple contrasts in Contrast-CAT on attribution quality, measured by average AUC for AOPC and LOdds across datasets under the MoRF setting.

The AOPC metric shows a sharp improvement as the number of references increases from 0 to 5. After 5 references, the AUC continues to increase, stabilizing between 25 and 30 references. In contrast, the LOdds metric exhibits a sharp decline as the number of references increases, starting at approximately 0.30 and dropping steadily, stabilizing around 0.10 after 10 references and reaching its minimum at 30 references. These results indicate that more references improve attribution quality, with the best performance at 30, which we use in our experiments.

5.7 THE EFFECT OF CONTRASTING REFERENCES

Table 4 presents the impact of the parameter γ in the condition for selecting contrastive references, $f_c(r) < \gamma$, on Contrast-CAT's attribution quality. This condition ensures

Table 2: The effect of our activation contrasting approach, measured under the MoRF (Most Relevant First) setting. 'Random' uses randomly selected references (the mean values over 30 repetitions are reported), 'Same' uses references from the same class as the target, and 'Contrasting' refers to the suggested Contrast-CAT. The best results are in boldface.

Dataset	Amazon		Yelp		SST2		IMDB	
Reference	AOPC↑	LOdds↓	AOPC↑	LOdds↓	AOPC↑	LOdds↓	AOPC↑	LOdds↓
Random	0.513	0.306	0.496	0.323	0.433	0.398	0.634	0.213
Same	0.144	0.667	0.159	0.650	0.089	0.728	0.124	0.614
Contrasting	**0.703**	**0.117**	**0.687**	**0.131**	**0.654**	**0.157**	**0.738**	**0.101**

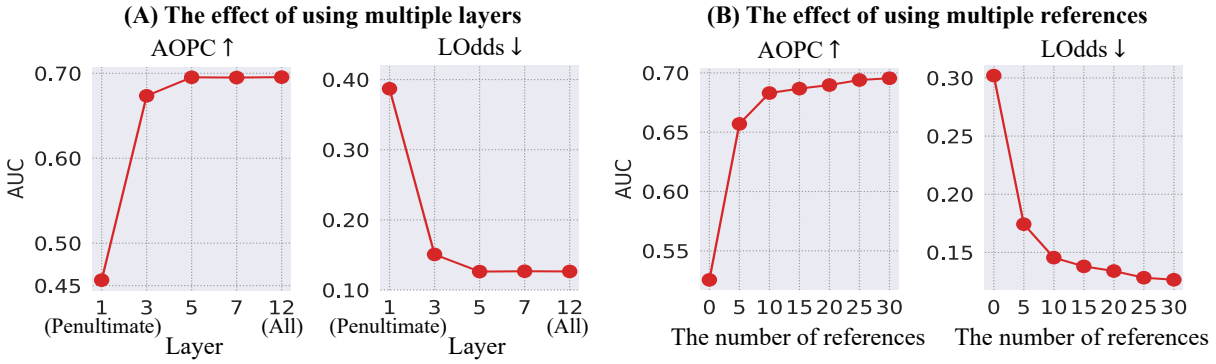

Figure 5: Comparison of Contrast-CAT's attribution quality measured under the MoRF (Most Relevant First) setting: (A) varying the number of layers from penultimate to all, and (B) varying the number of reference samples from 0 to 30.

Table 3: The results of the confidence evaluation, showing averaged rank correlation values. The values below 0.05 (marked in gray) indicate that attributions tend to be class-distinct, as desired.

Method	Dataset			
	Amazon	Yelp	SST2	IMDB
RawAtt	1.00	1.00	1.00	1.00
Rollout	1.00	1.00	1.00	1.00
Att-grads	< 0.05	< 0.05	< 0.05	< 0.05
Att×Att-grads	< 0.05	< 0.05	< 0.05	< 0.05
Grad-SAM	0.158	0.138	0.282	0.084
Full LRP	0.732	0.629	0.712	0.533
Partial LRP	0.952	0.924	0.957	0.859
TransAtt	0.153	0.135	0.342	0.061
CAT	< 0.05	< 0.05	< 0.05	< 0.05
AttCAT	< 0.05	< 0.05	< 0.05	< 0.05
TIS	< 0.05	< 0.05	< 0.05	< 0.05
Contrast-CAT	< 0.05	< 0.05	< 0.05	< 0.05

that selected reference activations do not strongly respond to the target class c, thereby helping to reduce non-target signals within the target activation by contrasting it with the selected reference activations.

We evaluated Contrast-CAT's faithfulness by varying γ from 0.1 to 0.001, and reported average AUC values for AOPC and LOdds across datasets under the MoRF setting. The

Table 4: Impact of the parameter γ in the condition $f_c(r) < \gamma$ on the attribution quality of Contrast-CAT.

γ	0.1	0.01	0.001
AOPC↑	0.627	0.651	0.696
LOdds↓	0.450	0.448	0.127

results in Table 4 indicate that a smaller γ improves Contrast-CAT's attribution quality, highlighting the benefits of low-activation references for activation contrasting, as described in Section 4.2.

6 CONCLUSION

In this work, we introduced Contrast-CAT, a novel activation-based attribution method that leverages activation contrasting to generate high-quality token-level attribution map. Our extensive experiments demonstrated that Contrast-CAT significantly outperforms state-of-the-art methods across various datasets and models.

Despite its effectiveness, Contrast-CAT requires reference points whose activations will be available during the creation of attribution maps. While we minimized overhead with a pre-built reference library, its storage requirements grow with the number of classes and activation size. Future work will explore lower-cost alternative tensors.

As the demand for interpretable AI grows to support safety, security, and trustworthiness, we believe Contrast-CAT represents a meaningful step toward improving the transparency of transformer-based models.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2024-00439819, AI-Based Automated Vulnerability Detection and Safe Code Generation) and by the IITP-ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT)(IITP-2025-RS-2020-II201749).

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, 2020.

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

Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. Grad-sam: Explaining transformers via gradient self-attention maps. In *CIKM*, 2021.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. In *ACL*, 2020.

George Chrysostomou and Nikolaos Aletras. Enjoy the salience: Towards better transformer-based faithful explanations with word salience. In *EMNLP*, 2021.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? An analysis of BERT's attention. In *ACL (workshop)*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. Visualizing and understanding neural machine translation. In *ACL*, 2017.

Jesse Dunietz, Elham Tabassi, Mark Latonero, and Kamie Roberts. A plan for global engagement on ai standards. NIST Trustworthy and Responsible AI, National Institute of Standards and Technology, Gaithersburg, MD, 2024. URL `https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=958389`.

Alexandre Englebert, Sédrick Stassin, Géraldin Nanfack, Sidi Ahmed Mahmoudi, Xavier Siebert, Olivier Cornu, and Christophe De Vleeschouwer. Explaining through transformer input sampling. In *ICCV (workshop)*, 2023.

European Commission. Artificial intelligence act (regulation (eu) 2024/1689). Official Journal of the European Union, 2024. URL `https://artificialintelligenceact.eu/the-act/`.

Jindong Gu, Yinchong Yang, and Volker Tresp. Understanding individual decisions of cnns via contrastive backpropagation. In *ACCV*, 2018.

Sungmin Han, Jeonghyun Lee, and Sangkyun Lee. Activation fine-tuning of convolutional neural networks for improved input attribution based on class activation maps. *Applied Sciences*, 2022.

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

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *ACL*, 2019.

Maurice George Kendall. *Rank correlation methods.* C. Griffin, 1955.

Sangkyun Lee and Sungmin Han. Libra-CAM: an activation-based attribution based on the linear approximation of deep neural nets and threshold calibration. In *IJCAI*, 2022.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL-HLT*, 2011.

André F. T. Martins and Ramón F. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *ICML*, 2016.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *NAACL*, 2022.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. Quantifying context mixing in transformers. In *EACL*, 2023.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. *Layer-wise relevance propagation: an overview*, pages 193–209. Springer International Publishing, 2019.

Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *NAACL-HLT*, 2018.

Damian Pascual, Gino Brunner, and Roger Wattenhofer. Telling BERT's full story: from local attention to global aggregation. In *EACL*, 2021.

V Petsiuk. Rise: randomized input sampling for explanation of black-box models. In *BMVC*, 2018.

Yao Qiang, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu. AttCAT: explaining transformers via attentive class activation tokens. In *NIPS*, 2022.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *ICML*, 2022.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

The White House. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House Presidential Actions, 2023.

Jacob Turton, Robert Elliott Smith, and David Vinson. Deriving contextualised semantic features from bert (and other transformer model) embeddings. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *ACL*, 2019.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: score-weighted visual explanations for convolutional neural networks. In *CVPR (workshop)*, 2020.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *EMNLP-IJCNLP*, 2019.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.