

Discretized Integrated Gradients for Explaining Language Models

Soumya Sanyal

University of Southern California
soumyasa@usc.edu

Xiang Ren

University of Southern California
xiangren@usc.edu

Abstract

As a prominent attribution-based explanation algorithm, Integrated Gradients (IG) is widely adopted due to its desirable explanation axioms and the ease of gradient computation. It measures feature importance by averaging the model’s output gradient interpolated along a *straight-line* path in the input data space. However, such straight-line interpolated points are not representative of text data due to the inherent discreteness of the word embedding space. This questions the faithfulness of the gradients computed at the interpolated points and consequently, the quality of the generated explanations. Here we propose Discretized Integrated Gradients (DIG), which allows effective attribution along non-linear interpolation paths. We develop two interpolation strategies for the discrete word embedding space that generates interpolation points that lie close to actual words in the embedding space, yielding more faithful gradient computation. We demonstrate the effectiveness of DIG over IG through experimental and human evaluations on multiple sentiment classification datasets. We provide the source code of DIG to encourage reproducible research ¹.

1 Introduction

In the past few years, natural language processing has seen tremendous progress, largely due to strong performances yielded by pre-trained language models (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020). But even with this impressive performance, it can still be difficult to understand the underlying reasoning for the preferred predictions leading to distrust among end-users (Lipton, 2018). Hence, improving model interpretability has become a central focus in the community with an increasing effort in developing methods that can explain model behaviors (Ribeiro et al., 2016; Binder et al., 2016; Li et al., 2016; Sundararajan et al.,

Input: the movie was **good** !

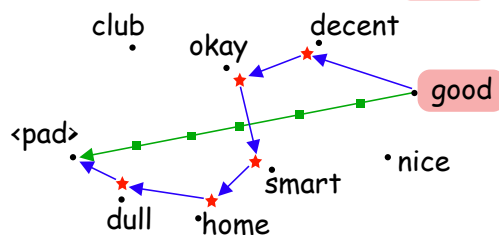


Figure 1: **An illustration of paths used in IG and DIG.** IG uses a straight line interpolation with points as depicted by green squares. In contrast, DIG uses a non-linear path (shown in blue) with interpolation points (red stars) lying close to words in the embedding space.

2017; Shrikumar et al., 2017; Lundberg and Lee, 2017; Murdoch et al., 2018).

Explanations in NLP are typically represented at a word-level or phrase-level by quantifying the contributions of the words or phrases to the model’s prediction by a scalar score. These explanation methods are commonly referred as *attribution-based methods* (Murdoch et al., 2018; Ancona et al., 2018). Integrated Gradients (IG) (Sundararajan et al., 2017) is a prominent attribution-based explanation method used due to the many desirable explanation axioms and ease of gradient computation. It computes the partial derivatives of the model output with respect to each input feature as the features are interpolated along a straight-line path from the given input to a baseline value. For example, say we want to compute the attribution for the word “good” in the sentence “the movie was good!” using IG. The straight-line interpolation path used by IG is depicted in green in Figure 1. Here, the baseline word is defined as the “<pad>” embedding and the green squares are the intermediate interpolation points in the embedding space.

While this method can be used for attributing inputs in both continuous (e.g., image, audio, etc.) and discrete (e.g., text, molecules, etc.) domains (Sundararajan et al., 2017), their usage in the dis-

¹<https://github.com/INK-USC/DIG>

crete domain has some limitations. Since the interpolation is done along a straight-line path joining the input word embedding and the baseline embedding (“<pad>” in Figure 1), the interpolated points are not necessarily representative of the discrete word embedding distribution. Specifically, let a dummy word embedding space be defined by the words represented by black dots in Figure 1. Then we can see that some of the green squares can be very far-off from any original word in the embedding space. Since the underlying language model is trained to effectively work with the specific word embedding space as input, using these out-of-distribution green interpolated samples as intermediate inputs to calculate gradients can lead to sub-optimal attributions.

To mitigate these limitations, we propose a Discretized integrated gradients (DIG) formulation by relaxing the constraints of searching for interpolation points along a straight-line path. Relaxing this linear-path constraint leads to a new constraint on the interpolation paths in DIG that points along the path should be monotonically situated between the input word embedding and the baseline embedding. Hence, in DIG, our main objective is to monotonically interpolate between the input word embedding and baseline such that the intermediate points are close to real data samples. This would ensure that the interpolated points are more representative of the word embedding distribution, enabling more faithful model gradient computations. To this end, we propose two interpolation strategies that search for an optimal anchor word embedding in the real data space and then modify it such that it lies monotonically between the input word and baseline (see Fig. 1 for an illustration).

We apply DIG using our proposed interpolation algorithms to generate attributions for three pre-trained language models - BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), and RoBERTa (Liu et al., 2019), each fine-tuned separately on three sentiment classification datasets - SST2 (Socher et al., 2013), IMDB (Maas et al., 2011), and Rotten Tomatoes (Pang and Lee, 2005). We find that our proposed interpolation strategies achieve a superior performance compared to integrated gradients and other gradient-based baselines on eight out of the nine settings across different metrics. Further, we also observe that on average, end-users find explanations provided by DIG to be more plausible justifications of model behavior

than the explanations from other baselines.

2 Method

In this section, we first describe our proposed Discretized integrated gradients (DIG) and the desirable explanation axioms satisfied by it. Then we describe an interpolation algorithm that leverages our DIG in discrete textual domains. Please refer to Appendix A for a brief introduction of the attribution-based explanation setup and the integrated gradients method.

2.1 Discretized integrated gradients

Below, we define our DIG formulation that allows interpolations along non-linear paths:

$$\text{DIG}_i(x) = \int_{x_i^k=x_i'}^{x_i} \frac{\partial F(x^k)}{\partial x_i} dx_i^k. \quad (1)$$

Here, x_i^k refers to the i^{th} dimension of the k^{th} interpolated point between input x and baseline x' and F is a neural network. The only constraint on x_i^k 's is that each interpolation should be monotonic between x_i and x_i' , i.e., $\forall j, k \in \{1, \dots, m\}; j < k$,

$$\begin{aligned} x_i' \leq x_i^j \leq x_i^k \leq x_i & \text{ if } x_i' \leq x_i, \\ x_i' \geq x_i^j \geq x_i^k \geq x_i & \text{ otherwise.} \end{aligned} \quad (2)$$

Here m is the total number of steps for interpolation. This constraint is essential because it allows approximating the integral in Eq. 1 using Riemann summation² which requires monotonic paths. We note that the interpolation points used by IG naturally satisfy this constraint since they lie along a straight line joining x and x' . The key distinction of our formulation from IG is that DIG is agnostic of any fixed step size parameter α and thus allows non-linear interpolation paths in the embedding space. The integral approximation of DIG is defined as follows:

$$\text{DIG}_i^{\text{approx}}(x) = \sum_{k=1}^m \frac{\partial F(x^k)}{\partial x_i} \times (x_i^{k+1} - x_i^k), \quad (3)$$

where m is the total number of steps considered for the approximation.

2.2 Axioms satisfied by DIG

As described in prior works (Sundararajan et al., 2017; Shrikumar et al., 2017), a good explanation

²https://en.wikipedia.org/wiki/Riemann_sum

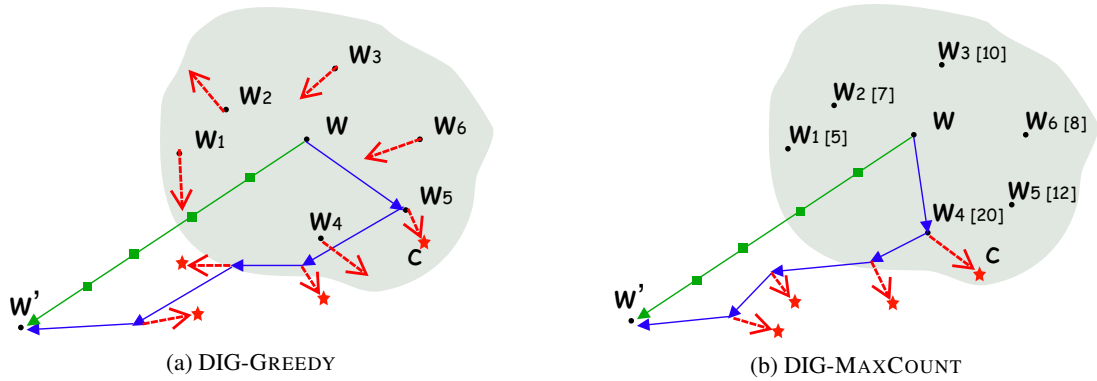


Figure 2: **Overview of paths used in DIG and IG.** The gray region is the neighborhood of w . Green line depicts the straight-line path used by IG. **Left:** In DIG-GREEDY, we first monotonize each word in the neighborhood (red arrow) and the word closest to its corresponding monotonic point is selected as the anchor (w_5 since the red arrow of w_5 has the smallest magnitude). **Right:** In DIG-MAXCOUNT we select the word with the highest number of monotonic dimensions (count shown in $[\cdot]$) as the anchor word (w_4), followed by changing the non-monotonic dimensions of w_4 (red arrow to c). Repeating this iteratively gives the non-linear blue path for DIG with the red stars as interpolation points. Please refer to Section 2.1 for more details. Figure best viewed in color.

algorithm should satisfy certain desirable axioms which justify the use of the algorithm for generating model explanations. Similar to IG, DIG also satisfies many such desirable axioms. First, DIG satisfies *Implementation Invariance* which states that attributions should be identical for two *functionally equivalent* models. Two models are functionally equivalent if they have the same output for the same input, irrespective of any differences in the model’s internal implementation design. Further, DIG satisfies *Completeness* which states that the sum of the attributions for an input should add up to the difference between the output of the model at the input and the baseline, i.e., $\sum_i \text{DIG}_i(x) = F(x) - F(x')$. This ensures that if $F(x') \approx 0$ then the output is completely attributed to the inputs. Thirdly, DIG satisfies *Sensitivity* which states that attributions of inputs should be zero if the model does not depend (mathematically) on the input. Please refer to Appendix B for further comparisons of DIG with IG.

2.3 Interpolation algorithm

Here, we describe our proposed interpolation algorithm that searches for intermediate interpolation points between the input word embedding and the baseline embedding. Once we have the desired interpolation points, we can use Equation 3 to compute the word attributions similar to the IG algorithm. Please refer to Section A.2 for more details about application of IG to text.

Design Consideration. First, we discuss the key design considerations we need to consider of our interpolation algorithm. Clearly, our interpolation points need to satisfy the monotonicity constraints defined in Equation 2 so that we can use the Riemann sum approximation of DIG. Hence, we need to ensure that every intermediate point lies in a monotonic path. Also, the interpolation points should lie close to the original words in the embedding space to ensure that the model gradients faithfully define the model behavior.

Now, we define the notion of *closeness* for our specific use-case of explaining textual models. To calculate how far the interpolated words are from some true word embedding in the vocabulary, we can compute the distance of the interpolated point from the nearest word in the vocabulary. We define this as the word-approximation error (WAE). More specifically, if w^k denotes the k^{th} interpolation point for a word w , then its word-approximation error along the interpolated path is defined as:

$$\text{WAE}_w = \frac{1}{m} \sum_{k=1}^m \min_{x \in V} \text{dist}(w^k - x), \quad (4)$$

where V is the embedding matrix of all the words in the vocabulary. WAE of a sentence is the average WAE of all words in the sentence. Intuitively, minimizing WAE will ensure that the interpolated points are close to some real word embedding in the vocabulary which in turn ensures that output gradients of F are not computed for some out-of-distribution unseen embedding points.

We observe that to minimize WAE without the monotonic constraints defined in Section 2.1, one can define some heuristic to search for interpolation points that belong to the set V (i.e., select words from the vocabulary as interpolation points), leading to a zero WAE. Motivated by this, for a given input word embedding, we first search for an anchor word from the vocabulary that can be considered as the next interpolation point. Since the anchor point need not be monotonic w.r.t. the given input, we then optimally perturb the dimensions of the anchor word so that they satisfy the monotonicity constraints in Equation 2. This perturbed point becomes our first interpolation. For subsequent interpolation points, we repeat the above steps using the previous anchor and perturbed points. Formally, we break our interpolation algorithm into two parts:

- (i) **ANCHORSEARCH**: In this step, given the initial word embedding w , we search for an anchor word embedding $a \in V$.
- (ii) **MONOTONIZE**: This step takes the anchor embedding a and modifies its dimensions to create a new embedding c such that all dimensions of c are monotonic between the input w and the baseline w' .

Overall, given an initial input word embedding w and a baseline embedding w' , our interpolation algorithm interpolates points from w to w' (which is in decreasing order of k in Eq. 3). It proceeds by calling **ANCHORSEARCH** on w to get an anchor word a . Then, it applies **MONOTONIZE** on a to get the monotonic embedding c . This is our first interpolated point (in reverse order), i.e., $c = w^{m-1}$. Now, the a becomes the new w for the next iteration and the process continues till m steps. Next, we describe in detail our specific formulations of the **MONOTONIZE** and **ANCHORSEARCH** algorithms.

MONOTONIZE: In this step, given an anchor word embedding a , we modify the non-monotonic dimensions of a such that they become monotonic w.r.t. w and w' . The monotonic dimensions of a vector a is given by:

$$M_a = \{j \mid w'_j \leq a_j \leq w_j, j \in \{1, \dots, D\}\} \\ \cup \{j \mid w'_j \geq a_j \geq w_j, j \in \{1, \dots, D\}\},$$

where D is the word embedding dimension. The number of monotonic dimensions is given by

the size of the set defined as $|M_a|$. Thus, the non-monotonic dimensions \overline{M}_a is the set complement of the monotonic dimensions, i.e., $\overline{M}_a = \{1, \dots, D\} - M_a$, where the subtraction is the set-diff operation. Let the final monotonic vector be c . We define the **MONOTONIZE** operations as follows:

$$c[M_a] \leftarrow a[M_a], \\ c[\overline{M}_a] \leftarrow w[\overline{M}_a] - \frac{1}{m} \times (w[\overline{M}_a] - w'[\overline{M}_a]),$$

where m is the total number of interpolation points we want to select in the path. It can be easily seen that c is monotonic w.r.t. w and w' according to the definition in Equation 2.

ANCHORSEARCH: First, we preprocess the word embedding in V to find the top- K nearest neighbor for each word. We consider this neighborhood for candidate anchor selection. Let us denote the K -neighbors for a word w by $KNN_V(w)$. We define two heuristics to search for the next anchor word: **GREEDY** and **MAXCOUNT**.

In the **GREEDY** heuristic, we first compute the monotonic embedding corresponding to each word in the neighborhood $KNN_V(w)$ using the **MONOTONIZE** step. Then, we select the anchor word a that is closest to its corresponding monotonic embedding obtained from the above step. This can be thought of as minimizing the WAE metric for a single interpolated word. The key intuition here is to locally optimize for smallest perturbations at each iterative selection step. This heuristic is depicted in Figure 2a and the algorithm is presented in Algorithm 1 in Appendix.

In the **MAXCOUNT** heuristic, we select the anchor a as the word in $KNN_V(w)$ with the highest number of monotonic dimensions. Precisely, the anchor is given by:

$$a = \arg \max_{a' \in KNN_V(w)} |M_{a'}|.$$

The intuition of this heuristic is that the vector with highest number of monotonic dimensions would require the minimum number of dimensions being perturbed in the **MONOTONIZE** step and hence, would be close to a word in the vocabulary. This heuristic is depicted in Figure 2b and the algorithm is presented in Algorithm 2 in Appendix.

3 Experimental Setup

In this section, we describe the datasets and models used for evaluating our proposed algorithm.

Method	DistilBERT				RoBERTa				BERT			
	LO ↓	Comp ↑	Suff ↓	WAE ↓	LO ↓	Comp ↑	Suff ↓	WAE ↓	LO ↓	Comp ↑	Suff ↓	WAE ↓
Grad*Inp	-0.402	0.112	0.375	-	-0.318	0.085	0.398	-	-0.502	0.168	0.366	-
DeepLIFT	-0.196	0.053	0.489	-	-0.300	0.078	0.432	-	-0.175	0.063	0.470	-
GradShap	-0.778	0.216	0.308	-	-0.523	0.168	0.347	-	-0.686	0.225	0.333	-
IG	-0.950	0.248	0.275	0.344	-0.738	0.222	0.250	0.669	-0.670	0.237	0.396	0.302
DIG-GREEDY	-1.222	0.310	0.237	0.229	-0.756	0.218	0.215	0.460	-0.879	0.292	0.374	0.249
DIG-MAXCOUNT	-1.259	0.307	0.241	0.227	-0.826	0.227	0.238	0.439	-0.777	0.272	0.377	0.173

Table 1: Comparison of variants of DIG with baselines on three LMs fine-tuned on SST2 dataset. ‘-’ denotes that the WAE metric is not computable for that setting. We observe that DIG outperforms the baselines on all three LMs. Please refer to Section 4.1 for more details.

Method	DistilBERT				RoBERTa				BERT			
	LO ↓	Comp ↑	Suff ↓	WAE ↓	LO ↓	Comp ↑	Suff ↓	WAE ↓	LO ↓	Comp ↑	Suff ↓	WAE ↓
Grad*Inp	-0.197	0.081	0.212	-	-0.195	0.043	0.279	-	-0.731	0.102	0.231	-
DeepLIFT	-0.021	-0.009	0.534	-	-0.157	0.028	0.340	-	-0.335	0.023	0.486	-
GradShap	-0.473	0.185	0.154	-	-0.416	0.129	0.196	-	-0.853	0.190	0.255	-
IG	-0.446	0.182	0.224	0.379	-0.733	0.226	0.084	0.708	-0.641	0.107	0.295	0.333
DIG-GREEDY	-0.878	0.319	0.133	0.256	-0.683	0.198	0.100	0.484	-1.152	0.221	0.240	0.287
DIG-MAXCOUNT	-0.795	0.296	0.152	0.255	-0.470	0.121	0.213	0.470	-0.995	0.195	0.245	0.190

Table 2: Comparison of variants of DIG with baselines on three LMs fine-tuned on IMDB dataset. We observe that DIG outperforms the baselines on DistilBERT and BERT models. Please refer to Section 4.1 for more details.

Datasets. The SST2 (Socher et al., 2013) dataset has 6920/872/1821 example sentences in the train/dev/test sets. The task is binary classification into positive/negative sentiment. The IMDB (Maas et al., 2011) dataset has 25000/25000 example reviews in the train/test sets with similar binary labels for positive and negative sentiment. Similarly, the Rotten Tomatoes (RT) (Pang and Lee, 2005) dataset has 5331 positive and 5331 negative review sentences. We use the processed dataset made available by HuggingFace Dataset library³ (Wolf et al., 2020b).

Language Models. We use pre-trained BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), and RoBERTa (Liu et al., 2019) text classification models individually fine-tuned for SST2, IMDB, and RT datasets.⁴ The fine-tuned checkpoints used are provided by the HuggingFace library (Wolf et al., 2020a).

Evaluation Metrics. Following prior literature, we use the following three automated metrics:

- **Log-odds (LO)** score (Shrikumar et al., 2017) is defined as the average difference of the

³<https://github.com/huggingface/datasets>

⁴Note that the vocabulary matrix V that is used in DIG is the *word_embeddings* layer of the language models in HuggingFace (Wolf et al., 2020a).

negative logarithmic probabilities on the predicted class before and after masking the top $k\%$ words with zero padding. Lower scores are better.

- **Comprehensiveness (Comp)** score (DeYoung et al., 2020) is the average difference of the change in predicted class probability before and after removing the top $k\%$ words. Similar to Log-odds, this measures the influence of the top-attributed words on the model’s prediction. Higher scores are better.
- **Sufficiency (Suff)** score (DeYoung et al., 2020) is defined as the average difference of the change in predicted class probability before and after keeping only the top $k\%$ words. This measures the adequacy of the top $k\%$ attributions for model’s prediction.

Please refer to Appendix C for more details about the evaluation metrics. We use $k = 20\%$ in our experiments. In Appendix D we further analyze the effect of changing top- $k\%$ on the metrics. Additionally, we use our proposed word-approximation error (WAE) metric to compare DIG with IG.

Method	DistilBERT				RoBERTa				BERT			
	LO ↓	Comp ↑	Suff ↓	WAE ↓	LO ↓	Comp ↑	Suff ↓	WAE ↓	LO ↓	Comp ↑	Suff ↓	WAE ↓
Grad*Inp	-0.152	0.068	0.315	-	-0.158	0.054	0.406	-	-0.801	0.204	0.398	-
DeepLIFT	-0.077	0.017	0.372	-	-0.150	0.050	0.413	-	-0.388	0.096	0.438	-
GradShap	-0.298	0.156	0.270	-	-0.290	0.128	0.338	-	-0.809	0.235	0.388	-
IG	-0.424	0.208	0.189	0.348	-0.368	0.149	0.317	0.677	-0.789	0.203	0.418	0.305
DIG-GREEDY	-0.501	0.257	0.184	0.329	-0.393	0.148	0.294	0.465	-1.056	0.267	0.416	0.251
DIG-MAXCOUNT	-0.467	0.231	0.190	0.230	-0.361	0.133	0.332	0.444	-0.874	0.237	0.430	0.178

Table 3: Comparison of variants of DIG with baselines on three LMs fine-tuned on Rotten Tomatoes dataset. We observe that DIG outperforms the baselines on all three LMs. Please refer to Section 4.1 for more details.

4 Results

4.1 Performance Comparison

We compare DIG with four representative gradient-based explanation methods - Gradient*Input (Grad*Inp) (Shrikumar et al., 2016), DeepLIFT (Shrikumar et al., 2017), GradShap (Lundberg and Lee, 2017), and integrated gradients (Sundararajan et al., 2017). For the IMDB and RT datasets, we randomly sample a subset of 2,000 reviews from the public test sets to compare the different methods, due to computation costs. For the SST2 dataset, we use the complete set of 1,821 test sentences. The results are shown in Tables 1, 2, and 3 for SST2, IMDB, and Rotten Tomatoes respectively.

Comparison with baselines. First, we observe that across the nine different settings we studied (three language models per dataset), DIG consistently outperforms the baselines on eight of the settings. This is valid for all the metrics. We also note that the WAE metric is lower for all variants of DIG compared to IG. This validates that our proposed interpolation strategies for DIG is able to considerably reduce the word-approximation error in the interpolated paths and consistently improving performance on all three explanation evaluation metrics considered.

Comparison between variants of DIG. Second, we observe that on average, DIG-GREEDY performs better than DIG-MAXCOUNT. Specifically, we find that DIG-MAXCOUNT doesn’t outperform DIG-GREEDY by significantly large margins on any setting (while the opposite is true for one setting - RoBERTa fine-tuned on IMDB dataset). This could be because the DIG-GREEDY strategy ensures that the monotonic point c is always close to the anchor a due to the locally greedy selection at each step which is not explicitly guaranteed by DIG-MAXCOUNT. But overall, we do not find any

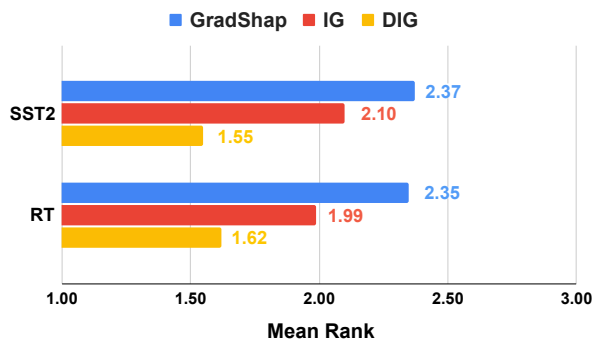


Figure 3: Result of human evaluation on DistilBERT model fine-tuned on SST2 dataset and BERT model fine-tuned on Rotten Tomatoes dataset. A lower mean rank means higher trustworthy explanation algorithm. For more details, refer to Section 4.2

specific performance trend between the two proposed variants and plan to study the influence of the embedding distribution in future works.

Analysis. Finally, though we are able to achieve good reductions in WAE, we note that the WAE for our interpolation algorithms are not close to zero yet. This leaves some scope to design better interpolation algorithms in future. Moreover, we find that the average Pearson correlation between log-odds and WAE is 0.32 and the correlation is 0.45 if we consider the eight settings where we outperform IG. We discuss the correlations of all the settings in Appendix E. While this suggests a weak correlation between the two metrics, it is hard to comment if there is a causality between the two. This is partially because we believe selection of interpolation points should also take the semantics of the perturbed sentences into consideration, which we don’t strongly enforce in our strategies. Hence, we think that constraining interpolations in a semantically meaningful way is a promising direction to explore.

Method	SST2	IMDB	RT
IG	-0.950	-0.446	-0.424
DIG-RANDOMANCHOR	-1.217 ± 0.024	-0.834 ± 0.021	-0.474 ± 0.003
DIG-RANDOMNEIGHBOR	-1.247 ± 0.013	-0.854 ± 0.015	-0.460 ± 0.010
DIG (best)	-1.259	-0.878	-0.501

Table 4: Comparison of DIG with two ablation variants - DIG-RANDOMANCHOR and DIG-RANDOMNEIGHBOR on the DistilBERT model. We report 5-seed average log-odds score for the randomized methods. Please refer to Section 4.3 for more details.

4.2 Human Evaluation

To further understand the impact of our algorithm on end users, we conduct human evaluations of explanations from our method and the two top baselines - IG and GradShap. We perform the study on the DistilBERT model fine-tuned on SST2 dataset and the BERT model fine-tuned on Rotten Tomatoes dataset. Further, we select the best variant of DIG on each dataset for explanation comparisons. First, we pick 50 sample sentences from each dataset with lengths between 5 and 25 words for easier visualizations. Then, we convert the attributions from each method into word highlights, whose intensity is determined by the magnitude of the attributions. Finally, we show the highlighted sentence and the model’s predicted label to the annotators and ask them to rank the explanations on a scale of 1-3, “1” being the most comprehensive explanation that best justifies the prediction.

Figure 3 shows the mean rank of each explanation algorithm across the two datasets. We find that DIG has a significantly lower mean rank compared to IG ($p < .001$ on both SST2 and Rotten Tomatoes⁵). Thus, we conclude that explanations generated by DIG are also trustworthy according to humans. Please refer to Appendix G for visualizations and discussion on explanations generated by our methods.

4.3 Performance Analysis

In this section, we report the ablation of ANCHORSEARCH and the effect of path density on DIG. Please refer to Appendix F for ablations on neighborhood size and discussions on computational complexity.

Ablation Study on ANCHORSEARCH. We ablate our methods with two random variants - DIG-RANDOMANCHOR and DIG-

⁵We compute the p -value using Wilcoxon signed-rank test.

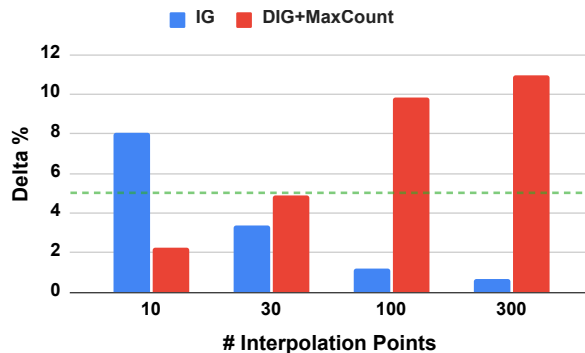


Figure 4: Effect of increasing number of interpolation points m on IG and DIG.

RANDOMNEIGHBOR, in which the ANCHORSEARCH step uses a random anchor selection heuristic. Specifically, in DIG-RANDOMANCHOR, the anchor is selected randomly from the complete vocabulary. Thus, this variant just ensures that the selected anchor is close to some word in the vocabulary which is not necessarily in the neighborhood. In contrast, the DIG-RANDOMNEIGHBOR selects the anchor randomly from the neighborhood without using our proposed heuristics MAXCOUNT or GREEDY. The log-odds metrics of IG, the two ablations, and our best variant of DIG for DistilBERT fine-tuned individually on all three datasets are reported in Table 4. We report 5-seed average for the randomized baselines. We observe that DIG-RANDOMANCHOR improves upon IG on all three datasets. This shows that generating interpolation points close to the words in the vocabulary improve the explanation quality. Further, we observe that DIG-RANDOMNEIGHBOR improves upon DIG-RANDOMANCHOR on log-odds metric. One reason could be that the words in a neighborhood are more semantically relevant to the original word, leading to more coherent perturbations for evaluating model gradients. Finally, we observe that, on average, our proposed method is better compared to selecting a random anchor in the neighborhood. This shows that our search strategies MAXCOUNT and GREEDY are indeed helpful.

Effect of Increasing Path Density. In integrated gradients, the *completeness* axiom (Section 2.2) is used to estimate if the integral approximation (Equation 6) error is low enough. This error is denoted as the Delta % error. If the error is high, users can increase the number of interpolation points m .

While DIG also satisfies the *completeness* axiom,

Factor f	Log-Odds ↓	WAE ↓	Delta % ↓
0	-1.259	0.227	4.926
1	-1.229	0.230	3.728
2	-1.184	0.232	2.752
3	-1.181	0.233	1.862

Table 5: Effect of up-sampling a path by a factor f on Delta % for DIG using $m = 30$.

error reduction by increasing m is infeasible. This is because increasing m in Equation 3 implicitly changes the integral path rather than increasing the density. Hence, to achieve an error reduction in DIG, we up-sample the interpolation path $P = \{w, w_1, w_2, \dots, w_{m-2}, w'\}$ with an up-sampling factor (f) of one as follows:

$$P_1 = \left\{w, \frac{w+w_1}{2}, w_1, \frac{w_1+w_2}{2}, \dots, \frac{w_{m-2}+w'}{2}, w'\right\},$$

i.e., we insert the mean of two consecutive points to the path. This essentially doubles the density of points in the path. Similarly, P_2 can be obtained by up-sampling P_1 , etc. DIG($m, f = 0$) refers to the standard DIG with no up-sampling.

Given that we have two hyperparameters m and f that determine the overall path density, we analyze the effect of each of these in Figure 4 and Table 5 respectively. The results are shown for DIG-MAXCOUNT applied on DistilBERT model finetuned on SST2 dataset. In Figure 4, we observe that as m increases, the Delta % of IG decreases as expected. But the trend is opposite for DIG. As discussed above, for DIG, the path length increases with increasing m , and hence, we attribute this trend to increasing difficulty in effectively approximating the integral for longer paths. Next, in Table 5, we observe that as the up-sampling factor f increases, the Delta % consistently decreases. We also find that our up-sampling strategy does not increase the WAE by a significant amount with increasing f , which is desirable. Thus, this confirms that our up-sampling strategy is a good substitute of increasing m for IG to effectively reduce the integral approximation error Delta %. Following Sundararajan et al. (2017), we choose a threshold of 5% average Delta to select the hyperparameters. For more discussions, please refer to Appendix F.1.

5 Related Works

There has been an increasing effort in developing interpretability algorithms that can help understand a neural network model’s behavior by explaining their predictions (Doshi-Velez and Kim,

2017; Gilpin et al., 2019). Attributions are a post-hoc explanation class where input features are quantified by scalar scores indicating the magnitude of contribution of the features toward the predicted label. Explanation algorithms that generate attributions can be broadly classified into two categories - model-agnostic algorithms, like LIME (Ribeiro et al., 2016), Input occlusion (Li et al., 2016), Integrated gradients ⁶(Sundararajan et al., 2017), SHAP (Lundberg and Lee, 2017), etc. and model-dependent algorithms, like LRP (Binder et al., 2016), DeepLIFT (Shrikumar et al., 2017), CD (Murdoch et al., 2018), ACD (Singh et al., 2019), SOC (Jin et al., 2020), etc. While the model-agnostic algorithms can be used as black-box explanation tools that can work for any neural network architecture, for the latter, one needs to understand the network’s architectural details to implement the explanation algorithm. Typically, model-dependent algorithms require specific layer decomposition rules (Ancona et al., 2018; Murdoch et al., 2018) which needs to be defined for all the components in the model. Model-agnostic methods usually work directly with the model outputs and gradients which are universally available.

Due to the many desirable explanation axioms and ease of gradient computation, there has been several extensions of integrated gradients. For example, Miglani et al. (2020) study the effect of saturation in the saliency maps generated by integrated gradients. Merrill et al. (2019) extend integrated gradients to certain classes of discontinuous functions in financial domains. Further, Jha et al. (2020) use KNNs and auto-encoders to learn latent paths for RNAs. Different from prior work, our focus here is to improve integrated gradients specifically for the discrete textual domain. While the idea of learning latent paths for text data is quite interesting, it brings a significant amount of challenge in successfully modeling such a complex latent space and hence, we leave this for future work.

6 Conclusion

In this paper, we proposed Discretized integrated gradients (DIG) which is effective in explaining models working with discrete text data. Further, we proposed two interpolation strategies - DIG-GREEDY and DIG-MAXCOUNT that generate non-

⁶Note that IG is strictly not a model-agnostic algorithm since it is defined for neural networks, but we still classify it as one since the scope of this work is limited to working on neural networks.

linear interpolation paths for word embedding space. Finally, we established the effectiveness of DIG over integrated gradients and other gradient-based baselines through experiments on multiple language models and datasets. We also conduct human evaluations and find that DIG enhances human trust on model predictions.

Acknowledgments

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600007, the DARPA MCS program under Contract No. N660011924033, the Defense Advanced Research Projects Agency with award W911NF-19-20271, NSF IIS 2048211, NSF SMA 1829268, USC Anenberg Graduate Fellowship, and gift awards from Google, Amazon, JP Morgan and Sony. We would like to thank all the collaborators in USC INK research lab for their constructive feedback on the work.

References

- Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. [Towards better understanding of gradient-based attribution methods for deep neural networks](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*. Springer.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- 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.
- Finale Doshi-Velez and Been Kim. 2017. [Towards a rigorous science of interpretable machine learning](#).
- Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2019. [Explaining explanations: An overview of interpretability of machine learning](#).
- Anupama Jha, Joseph K. Aicher, Matthew R. Gazzara, Deependra Singh, and Yoseph Barash. 2020. Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome Biology*, (1).
- Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. [Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. [Understanding neural networks through representation erasure](#). *ArXiv preprint*, abs/1612.08220.
- Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, (3).
- Yinhan Liu, Myle Ott, Naman Goyal, and Jingfei Du. 2019. [Roberta: A robustly optimized bert pre-training approach](#). *ArXiv preprint*, abs/1907.11692.
- Scott M. Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

- John Merrill, Geoff Ward, Sean Kamkar, Jay Budzik, and Douglas Merrill. 2019. Generalized integrated gradients: A practical method for explaining diverse ensembles.
- Vivek Miglani, Narine Kokhlikyan, Bilal Alsallakh, Miguel Martin, and Orion Reblitz-Richardson. 2020. Investigating saturation effects in integrated gradients.
- W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.
- Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.
- Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *ArXiv preprint*, abs/1605.01713.
- Chandan Singh, W. James Murdoch, and Bin Yu. 2019. Hierarchical interpretations for neural network predictions. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020a. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Thomas Wolf, Quentin Lhoest, Patrick von Platen, Yacine Jernite, Mariama Drame, Julien Plu, Julien Chaumond, Clement Delangue, Clara Ma, Abhishek Thakur, Suraj Patil, Joe Davison, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angie McMillan-Major, Simon Brandeis, Sylvain Gugger, François Lagunas, Lysandre Debut, Morgan Funtowicz, Anthony Moi, Sasha Rush, Philipp Schmid, Pierric Cistac, Victor Muštar, Jeff Boudier, and Anna Tordjmann. 2020b. Datasets. *GitHub. Note: <https://github.com/huggingface/datasets>*.

A Preliminaries

A.1 Attribution-based Explanations

Attribution-based explanations generate a scalar score for a given input feature that indicates the contribution (or importance) of that feature towards particular label (Ancona et al., 2018). Formally, let $x = [x_1, \dots, x_N] \in \mathbb{R}^N$ be an input to a model which produces an output $y = [y_1, \dots, y_C]$, where C is the total number of labels. For a given label (usually the label predicted by the model), attribution-based explanation methods compute the contribution $R^c = [R_1^c, \dots, R_N^c] \in \mathbb{R}^N$ of each feature.

A.2 Integrated gradients

Integrated gradients (IG) (Sundararajan et al., 2017) for an input x along the i^{th} dimension is defined as follows:

$$\text{IG}_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha. \quad (5)$$

Here, F is the neural network, x' is a baseline embedding, and α is the step size. Simply put, integrated gradients algorithm works by sampling points at a uniform spacing along a straight-line between the input and the baseline, and summing the model's gradient at the inputs for each interpolated points. To compute this integral efficiently, the authors propose a Riemann summation approximation defined below:

$$\text{IG}_i^{\text{approx}}(x) = (x_i - x'_i) \times \sum_{k=1}^m \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}, \quad (6)$$

where m is the total number of steps considered for the approximation.

Next, we briefly describe how IG is used to explain a model's prediction which takes a sentence as input (for example, the model can be a text classification network). Let $S = [w_0..w_n]$ be a sentence of length n and w_i be the i^{th} word embedding of the sentence. Also, let F be a text-classification model, i.e., $y = F(S)$. Then, IG calculates the attribution for each dimension of a word embedding w_i . The interpolation points required for Equation 6 are generated by linearly interpolating the word embedding between w_i and a baseline word embedding (usually chosen as the pad embedding). Then, using Eq. 6, the attribution for the i^{th} dimension of w is calculated. The final word attribution is the sum of the attributions for each dimension of the word embedding.

B Comparison with Integrated gradients and Path methods

It is easy to see that the approximation of integrated gradients is a special case of DIG. Note that the k^{th} linear interpolation of the i^{th} dimension of input x for IG can be represented as:

$$x_i^k = x'_i + \frac{k}{m} \times (x_i - x'_i). \quad (7)$$

Substituting Eq. 7 in Eq. 3 gives us Eq. 6.

Sundararajan et al. (2017) define path methods as the general form of integrated gradients that are applicable for all monotonic paths between the input and the baseline. Our DIG approach is a reformulation of the path method where the paths are not necessarily parameterized by α , making it more applicable for discrete data domain. Hence, DIG also satisfies all the theoretical properties applicable for path methods - Implementation Invariance, Sensitivity, Linearity, and Completeness. We refer the readers to Proposition 2 in Sundararajan et al. (2017) for more technical details.

C Evaluation Metrics

In this section, we redefine the evaluation metrics and state the formulations for each of them. In this work, we use the following three automated metrics:

- **Log-odds (LO) score** (Shrikumar et al., 2017) is defined as the average difference of the negative logarithmic probabilities on the predicted class before and after masking the top $k\%$ features with zero padding. Given the attribution scores generated by an explanation algorithm, we select the top $k\%$ words based on their attributions replace them with zero padding. More concretely, for a dataset with N sentences, it is defined as:

$$\log - \text{odds}(k) = \frac{1}{N} \sum_{i=1}^N \log \frac{p(\hat{y} | \mathbf{x}_i^{(k)})}{p(\hat{y} | \mathbf{x}_i)},$$

where \hat{y} is the predicted class, \mathbf{x}_i is the i^{th} sentence, and $\mathbf{x}_i^{(k)}$ is the modified sentence with top $k\%$ words replaced with zero padding. Lower scores are better.

- **Comprehensiveness (Comp) score** (DeYoung et al., 2020) is the average difference of the change in predicted class probability

Dataset	DistilBERT	RoBERTa	BERT
SST2	1.00	0.00	0.42
IMDB	0.98	-0.68	0.51
Rotten Tomatoes	0.21	0.22	0.23

Table 6: Pearson correlation between log-odds and WAE metrics for different dataset+LM settings. Please refer to Appendix E for more details.

before and after removing the top $k\%$ features. Similar to Log-odds, this measures the influence of the top-attributed words on the model’s prediction. It is defined as:

$$\text{Comp}(k) = \frac{1}{N} \sum_{i=1}^N p(\hat{y} | \mathbf{x}_i^{(k)}) - p(\hat{y} | \mathbf{x}_i).$$

Here $\mathbf{x}_i^{(k)}$ denotes the modified sentence with top $k\%$ words deleted from the sentence. Higher scores are better.

- **Sufficiency** (Suff) score (DeYoung et al., 2020) is defined as the average difference of the change in predicted class probability before and after keeping only the top $k\%$ features. This measures the adequacy of the top $k\%$ attributions for model’s prediction. It is defined in a similar fashion as comprehensiveness, except the $\mathbf{x}_i^{(k)}$ is defined as the sentence containing only the top $k\%$ words. Lower scores are better.

D Effect of top-k in evaluation metrics

In Figure 5, we visualize the effect of changing top-k% on log-odds, comprehensiveness, and sufficiency metrics for DistilBERT model fine-tuned on the SST2 dataset. We compare the two variants of our method: DIG-GREEDY and DIG-MAXCOUNT with Integrated Gradients. We observe that our method outperforms IG for all values of k. Specifically, we note that the gap between DIG and IG is initially non-existent but then gradually increases with increasing k in Figure 5 (a) and eventually saturates. This shows that although IG might be equally good as DIG at finding the top-5% important words, the explanations from IG are significantly misaligned from true model behavior for higher top-k values.

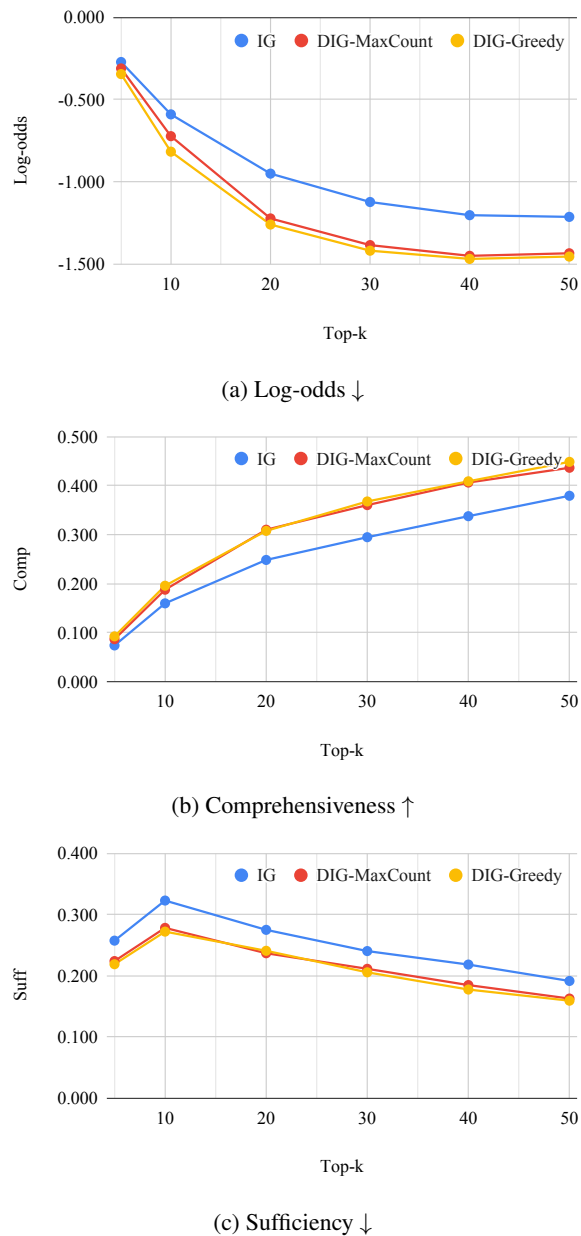


Figure 5: Effect of changing top-k% in log-odds, comprehensiveness, and sufficiency metric for the DistilBERT model fine-tuned on SST2 dataset.

m	IG		DIG	
	Log-Odds ↓	Delta % ↓	Log-Odds ↓	Delta % ↓
10	-0.984	8.064	-1.252	2.263
30	-0.950	3.394	-1.259	4.926
100	-0.933	1.235	-1.258	9.849
300	-0.940	0.703	-1.242	10.955

Table 7: Effect of increasing number of interpolation points m on Delta % for IG and DIG. Please refer to Appendix F.1 for more details.

Up-sampling factor f	Log-Odds ↓	WAE ↓	Delta % ↓
DIG ($m = 30, f = 0$)	-1.259	0.227	4.926
DIG ($m = 30, f = 1$)	-1.229	0.230	3.728
DIG ($m = 30, f = 2$)	-1.184	0.232	2.752
DIG ($m = 30, f = 3$)	-1.181	0.233	1.862

Table 8: Effect of up-sampling a path by a factor f on Delta % for DIG. For more details, refer to Appendix F.1.

E Correlation between Log-odds and WAE

We compute the Pearson correlation between log-odds and WAE for each dataset + LM pair. For this, we consider the metric values for IG, DIG-GREEDY, and DIG-MAXCOUNT and report the correlations for each setting in Table 6. We observe that, there is a strong correlation on average for DistilBERT. For BERT and RoBERTa we find a weak positive and negative correlation respectively.

F Ablation Studies

F.1 Effect of increasing path density

Here, we report the detailed analysis of the effect of increasing m and f in Tables 7 and 8 respectively. In Table 7, we report the Log-odds score along with Delta %. We do not note any consistent trend in Log-odds with increasing m for both IG and DIG. The results of IG suggest that, as long as the Delta % is *sufficiently* low, decreasing Delta % any further doesn't impact the explanations very significantly. Further, in Table 8, we report the WAE metrics to emphasize that our up-sampling strategy doesn't increase the WAE by a significant amount, which is desirable. Also, we note a consistent increase (although marginally) in Log-odds with decreasing Delta %. But per our previous observations on IG, we believe these changes do not imply a causal relation between the two.

K	Log-odds ↓	WAE ↓	Delta % ↓
10	-1.258	0.276	21.405
30	-1.263	0.310	12.228
100	-1.277	0.276	14.155
200	-1.194	0.295	10.647
300	-1.216	0.286	8.523
500	-1.259	0.227	4.926

Table 9: Effect of increasing the neighborhood size K of KNN_V for DIG. Please refer to Appendix F.2 for more details.

F.2 Effect of increasing neighborhood size

In this section, we study the effect of increasing the neighborhood size in DIG. The results are shown in Table 9. We observe a clear decreasing trend in Delta % with increasing neighborhood size, but there is no clear trend on Log-odds or WAE. Hence, we believe that the neighborhood size has little impact on the explanation quality, but we should still ensure sufficiently low Delta.

F.3 Discussion on computational complexity

In this section, we briefly discuss the computational complexity of our proposed interpolation strategies. The algorithms for DIG-GREEDY and DIG-MAXCOUNT are presented in Algorithms 1 and 2 respectively. From there, we observe that both our algorithms have a running time complexity of $\mathcal{O}(nmK)$, where n is the number of words, m is the number of interpolation points, and K is the KNN_V neighborhood size. While it is computationally feasible to parallelize the loops corresponding to n and K , the same cannot be said for the loop corresponding to m because we select the interpolation points iteratively. Although we empirically find in Section F.1 that a small number of interpolation points are sufficient to calculate the explanations, we believe this bottleneck can be further tackled through efficient design of non-iterative search algorithms. We leave this for future works.

G Visualizations of explanations

In this section, we present some interesting sentence visualizations based on explanations from DIG and IG for SST2 dataset in Figure 6. We show the sentence visualization and the model's predicted sentiment for the sentence for each explanation algorithm. In the visualizations, the red highlighted words denote positive attributions and

blue denotes negative attributions. That is, the explanation model suggests that the red highlighted words support the predicted label whereas the blue ones oppose (or undermine) the prediction. We observe that in many cases, DIG is able to highlight more plausible explanations. For example, in sentence pairs 1-7, clearly the DIG highlights are more inline with the model prediction. But we want to emphasize that it does not mean that our method always produces more plausible highlights. For example, for sentences 8-10, we observe that highlights from IG are more plausible than those of DIG. Hence, this shows that, while it could be a good exercise to visualize the attributions as a sanity check, we should rely more on automated metrics and human evaluations to correctly compare explanation algorithms.

Algorithm 1: DIG-GREEDY

Input : Sentence $s = [w_1, w_2, \dots, w_n]$,
 k -nearest neighbor graph for the
vocabulary KNN_V , number of
interpolation points m

Output : Interpolations

```

1  $points = []_{n*m}$ 
2 for  $i \leftarrow 1$  to  $n$  do
3   for  $j \leftarrow 1$  to  $m$  do
4      $dists = \{ \}$ 
5     for  $k \leftarrow 1$  to  $K$  do
6        $nbr \leftarrow KNN_V(w_i)[k]$ 
7        $c' \leftarrow \text{MONOTONIZE}(nbr, w_i)$ 
8        $dists[nbr] \leftarrow$   

        Distance( $nbr, c'$ )
9     end for
10     $a \leftarrow \arg \min_{a' \in dists} dists[a']$ 
11     $c \leftarrow \text{MONOTONIZE}(a, w_i)$ 
12     $points[i, j] \leftarrow c$ 
13  end for
14 end for
15 return  $points$ 

```

Algorithm 2: DIG-MAXCOUNT

Input : Sentence $s = [w_1, w_2, \dots, w_n]$,
 k -nearest neighbor graph for the
vocabulary KNN_V , number of
interpolation points m

Output : Interpolations

```

1  $points = []_{n*m}$ 
2 for  $i \leftarrow 1$  to  $n$  do
3   for  $j \leftarrow 1$  to  $m$  do
4      $a \leftarrow \arg \max_{a' \in KNN_V(w_i)} |M_{a'}|$ 
5      $c \leftarrow \text{MONOTONIZE}(a, w_i)$ 
6      $points[i, j] \leftarrow c$ 
7   end for
8 end for
9 return  $points$ 

```

DIG: [CLS] first **good** , then **bothersome** . [SEP] , positive
 IG: [CLS] first **good** , then **bothersome** . [SEP] , positive

DIG: [CLS] an **offbeat** , sometimes **gross** and **surprisingly** **appealing** **animated** **film** about the true meaning of the holidays . [SEP] , positive
 IG: [CLS] an **offbeat** , sometimes **gross** and **surprisingly** **appealing** **animated** **film** about the true meaning of the holidays . [SEP] , positive

DIG: [CLS] **do n ' t** let your festive **spirit** go this far . [SEP] , negative
 IG: [CLS] do n ' t **let** your festive spirit go this far . [SEP] , negative

DIG: [CLS] you have to pay attention to follow all the stories , but they ' re each **interesting** . [SEP] , positive
 IG: [CLS] you have to **pay** attention to follow all the stories , but they ' re each **interesting** . [SEP] , positive

DIG: [CLS] choose your reaction : a .) that sure is **funny** ! [SEP] , positive
 IG: [CLS] choose your reaction : a .) that sure is **funny** ! [SEP] , positive

DIG: [CLS] has a shambling **charm** . . . a **cheerfully** inconsequential **diversion** . [SEP] , positive
 IG: [CLS] **has** a **shambling** **charm** . . . a **cheerfully** inconsequential **diversion** . [SEP] , positive

DIG: [CLS] the movie ' s **ripe** , **enrapturing** **beauty** will tempt those willing to probe its **inscrutable** **mysteries** . [SEP] , positive
 IG: [CLS] the movie ' s **ripe** , enrapturing beauty **will tempt** those willing to probe its **inscrutable** **mysteries** . [SEP] , positive

DIG: [CLS] the spark of special anime magic here is **unmistakable** and **hard to resist** . [SEP] , positive
 IG: [CLS] the spark of special anime magic here is **unmistakable** and **hard to resist** . [SEP] , positive

DIG: [CLS] **even** with all those rough edges safely sanded down , the american **insomnia** is still **pretty darned** good . [SEP] , positive
 IG: [CLS] **even** with all those rough edges safely sanded down , the american insomnia is still **pretty darned** **good** . [SEP] , positive

DIG: [CLS] the issue of faith is not **explored** very deeply [SEP] , negative
 IG: [CLS] the issue of faith is **not explored** very deeply [SEP] , negative

Figure 6: Some example visualizations of attributions from DIG and IG for the DistilBERT model fine-tuned on SST2 dataset. The sentence visualization is followed by model’s sentiment prediction for the sentence. Here, the red highlighted words denote positive attributions and blue denotes negative attributions. For more details, please refer to Appendix G