ART: Attention-Regularized Transformers for Multi-Modal Robustness

Mohammed Bouri[1,3], **Mohammed Erradi**[1,2], **Adnane Saoud**[1]
[1]College of Computing, Mohammed VI Polytechnic University, Morocco,
[2]ENSIAS, University Mohamed V of Rabat, Morocco,
[3]CID Development, Morocco
mohammed.bouri@um6p.ma, mohammed.erradi@um6p.ma, adnane.saoud@um6p.ma,

Abstract

Transformers have become the standard in Natural Language Processing (NLP) and Computer Vision (CV) due to their strong performance, yet they remain highly sensitive to small input changes, often referred to as adversarial attacks, such as synonym swaps in text or pixel-level perturbations in images. These adversarial attacks can mislead predictions, while existing defenses are often domain-specific or lack formal robustness guarantees. We propose the *Attention-Regularized Transformer* (ART), a framework that enhances robustness across modalities. ART builds on the *Attention Sensitivity Tensor* (AST), which quantifies the effect of input perturbations on attention outputs. By incorporating an AST-based regularizer into training, ART encourages stable attention maps under adversarial perturbations in both text and image tasks. We evaluate ART on IMDB, QNLI, CIFAR-10, CIFAR-100, and Imagenette. Results show consistent robustness gains over strong baselines such as FreeLB and DSRM: up to $+36.9\%$ robust accuracy on IMDB and QNLI, and $+5$–25% on image benchmarks across multiple Vision Transformer (ViT) architectures, while maintaining or improving clean accuracy. ART is also highly efficient, training over $10\times$ faster than adversarial methods on text and requiring only $1.25\times$ the cost of standard training on images, compared to 1.5–$5.5\times$ for recent robust ViTs. Codes are available at https://github.com/cliclab-um6p/ART

1 Introduction

Transformer models have become increasingly popular in both Natural Language Processing (NLP) and Computer Vision (CV) tasks due to their strong performance and ability to capture complex dependencies in data (Devlin et al., 2019; Vaswani et al., 2017; Dosovitskiy et al., 2020). These models are built around a key idea called attention, which makes it possible to decide which parts of the input are most important when making a prediction. Despite their success, these models are known to be vulnerable to small changes in their input, known as adversarial attacks or adversarial perturbations. For instance, replacing a few words with synonyms in text, or changing some pixels slightly in images, can lead these models to produce incorrect predictions (Jin et al., 2020; Li et al., 2020). This vulnerability poses a significant risk when deploying Transformers in real-world scenarios, where reliability and robustness are crucial (Carlini and Wagner, 2017; Mao et al., 2022).

Recent research has explored adversarial training (Madry et al., 2017; Zhu et al., 2019) and input-space regularization (Li et al., 2021; Mao et al., 2022) to improve model robustness, though their effectiveness varies across tasks and settings. Despite these efforts, two main limitations remain. First, most methods are domain-specific, designed separately for either text or image tasks, making them hard to generalize. Second, they often lack formal theoretical guarantees, making it unclear why they improve robustness. In text classification, FreeLB++ (Li et al., 2021) enhances robustness through stronger embedding-space perturbations during adversarial training, while DSRM (Gao et al., 2023) models distributional shifts to estimate adversarial risk without explicit adversarial samples. These methods, however, remain text-specific and computationally expensive. In image classification, approaches such as SpecFormer (Hu et al., 2024) and LipsFormer (Qi et al., 2023) improve Vision Transformers (ViT) (Dosovitskiy et al., 2020) robustness by controlling global Lipschitz properties, e.g., penalizing large singular values or applying cosine normalization to attention. These techniques rely on loose global bounds that limit the model's ability to capture complex patterns.

To address these challenges, we introduce *Attention Sensitivity Tensor* (AST), a theoretical framework that quantifies how input perturbations affect attention outputs. Building on AST, we propose the *Attention-Regularized Transformer* (ART), a

Findings of the Association for Computational Linguistics: EACL 2026, pages 5510–5535
March 24-29, 2026

unified and theoretically grounded framework that enhances robustness across text and image domains by incorporating AST-based regularization during training, without altering model architecture.

Our main contributions are summarized as follows:

- **Unified Framework**: We formalize AST as a mathematical framework for analyzing attention sensitivity and propose ART, which leverages AST regularization to promote stable and robust attention patterns across text and image tasks.
- **Robustness Across Domains**: ART achieves state-of-the-art adversarial robustness. On NLP benchmarks (IMDB, QNLI), it improves robust accuracy by up to 36.9% while preserving clean accuracy; on vision datasets (CIFAR-10, CIFAR-100, Imagenette), it outperforms strong baselines by up to 29% under common attacks.
- **Efficiency and Practicality**: ART dramatically reduces computational overhead, training $> 10\times$ faster than adversarial methods on text tasks and requiring only $1.25\times$ the cost of standard training on images.

2 Related Work

2.1 Transformer robustness in text classification

2.1.1 Adversarial Training.

Adversarial training improves robustness by including perturbed examples in training. PGD-based methods (Madry et al., 2017) and FreeLB (Zhu et al., 2019) generate embedding-space perturbations via multi-step gradient updates. TA-VAT (Li and Qiu, 2020) extends this idea with token-level perturbations, enhancing interpretability, while FreeLB++ (Li et al., 2021) increases perturbation strength and steps, yielding stronger robustness with minimal loss of clean accuracy. Despite their effectiveness, these methods are computationally costly and require careful tuning.

2.1.2 Certified Defenses.

Certified defenses aim to provide provable robustness guarantees against perturbations. A widely used approach is *randomized smoothing*, which adds noise to the input and averages predictions over the perturbed samples. In NLP, this has been applied via synonym substitutions or noise added to embeddings (Zhang et al., 2024; Ye et al., 2020; Zeng et al., 2021). However, these methods often assume prior knowledge of the attacker's synonym choices, which is unrealistic in practice.

2.1.3 Regularization-Based Defenses.

Regularization-based methods enhance stability by encouraging consistent predictions under small input changes. InfoBERT (Wang et al., 2020) guides models to focus on important features, while DNE (Zhou et al., 2020) and ASCC (Dong et al., 2021) enforce consistency across semantically equivalent sentences, such as those modified with synonyms. These approaches are simple to apply, but generally less effective than adversarial training against strong attacks (Li et al., 2021).

2.2 Transformer robustness in image classification

2.2.1 Adversarial Training.

This approach improves robustness by training models on perturbed images. Attacks such as FGSM (Goodfellow et al., 2014) and PGD (Madry et al., 2017) generate pixel-level noise to mislead predictions, and ViTs trained with these perturbations gain resistance to such attacks. PGD-based training is a strong baseline, but as in text classification, adversarial training is computationally costly and may reduce clean accuracy if not carefully tuned.

2.2.2 Regularization-Based Defenses.

These methods improve stability without generating adversarial samples. SpecFormer (Hu et al., 2024) enhances robustness by penalizing large singular values, while LipsFormer (Qi et al., 2023) applies cosine normalization to stabilize attention. Other variants, such as L2Former (Kim et al., 2021) and LNFormer (Dasoulas et al., 2021), impose Lipschitz constraints in different ways. Although supported by theory, these methods rely on a global Lipschitz constant that measures sensitivity independent of direction. This scalar bound is often loose and conservative, particularly for Transformers where sensitivity varies across input dimensions or positions, failing to capture coordinate-wise variations and thereby limiting robustness precision and model expressiveness.

3 Preliminaries

We consider a classification task where an input X with label y is drawn from a dataset $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$. Let $f : X \to \hat{y}$ denote a

Transformer-based classifier that maps input X to predicted label $\hat{y}$.

In *text classification*, $X = \langle w_1, \ldots, w_N \rangle$ is a sequence of N tokens from a vocabulary $\mathcal{V}$, where N is the maximum sequence length. Transformers encode X via token and positional embeddings, followed by self-attention. In the context of text classification, adversarial attacks are generated via synonym substitutions: each token w_i has a set $\mathcal{S}(w_i)$ of k nearest neighbors within Euclidean distance d_e in embedding space, with the adversarial set defined as:

$$\mathcal{S}_{adv}(X) = \{\langle w'_1, \ldots, w'_N \rangle \mid w'_i \in \mathcal{S}(w_i) \cup \{w_i\}\}.$$

In *image classification*, where inputs X are images. Vision Transformers (ViTs) divide each image into N non-overlapping patches, embed them with positional information, and apply self-attention. These models are commonly attacked using a bounded additive perturbations δ under a norm constraint, with the adversarial set defined as:

$$\mathcal{I}_{adv}(X) = \{X' \mid \|X' - X\| \leq \delta\}.$$

Our goal is to design attention mechanisms that preserve predictions under both synonym-based substitutions in text and bounded additive perturbations in images:

$$\forall X' \in \mathcal{S}_{adv}(X) \cup \mathcal{I}_{adv}(X),\ f(X') = f(X) = \hat{y} = y.$$

To this end, we propose a unified regularization framework to enhance Transformer robustness across text and image domains.

4 ART: Theory and Framework

In this section, we present our main contribution, the Attention-Regularized Transformer (ART). We propose a theoretical concept, the Attention Sensitivity Tensor (AST), designed explicitly to quantify and control the sensitivity of self-attention mechanisms within Transformer models to input perturbations. We unify this concept across both text and image inputs, establishing a rigorous theoretical foundation for enhancing Transformers robustness.

4.1 Attention Sensitivity Tensor (AST)

The Attention Sensitivity Tensor (AST) explicitly quantifies the sensitivity of a Transformer's attention mechanism to input variations. By controlling AST values, we can directly enhance robustness against adversarial perturbations.

4.1.1 Self-Attention Mechanism

Self-attention is a fundamental operation in Transformer-based architectures, enabling each element in a sequence of text tokens or a grid of image patches to interact with every other element, and compute contextualized representations. As introduced by (Vaswani et al., 2017), self-attention operates over an input matrix $X \in \mathcal{X} \subseteq \mathbb{R}^{N \times d}$, where $\mathcal{X}$ is the set of possible inputs (words or images), N is the number of tokens in text or the number of patches in images, and d is the embedding size. The self-attention mechanism maps the input matrix X to an output in $\mathbb{R}^{N \times d_v}$ via the following mapping:

$$\begin{aligned} \mathcal{Z} : \mathcal{X} &\to \mathbb{R}^{N \times d_v} \\ X &\mapsto \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \end{aligned} \tag{1}$$

where the softmax function is applied row-wise to produce attention weights. The query, key, and value matrices Q, K, V are obtained through learned linear projections of the input, where $Q = XW_Q, K = XW_K, V = XW_V$ with $W_Q, W_K \in \mathbb{R}^{d \times d_k}$, and $W_V \in \mathbb{R}^{d \times d_v}$. Here, d_k and d_v denote the dimensions of the key/query and value vectors, respectively.

4.1.2 Definition of Attention Sensitivity Tensor (AST):

Consider the self-attention mapping $\mathcal{Z} : \mathcal{X} \to \mathbb{R}^{N \times d_v}$ defined in Eq. 1. A tensor $\mathcal{A} \in \mathbb{R}^{N \times d_v \times N \times d}$ is said to be an *Attention Sensitivity Tensor (AST)* of $\mathcal{Z}$, if the following condition holds: For all $(i, f, j, g) \in \mathcal{I}$, and for all $X \in \mathcal{X}$

$$\left\| \frac{\partial \mathcal{Z}[i, f]}{\partial X[j, g]} \right\| \leq \mathcal{A}[i, f, j, g], \tag{2}$$

where $\mathcal{I} = \{1, \ldots, N\} \times \{1, \ldots, d_v\} \times \{1, \ldots, N\} \times \{1, \ldots, d\}$, and $\mathcal{A}[i, f, j, g] \in \mathbb{R}$ denotes an upper bound on the partial derivative of the output feature f at position i with respect to the input feature g at position j.

The AST quantifies the sensitivity of each attention output dimension to input perturbations, providing a principled framework to analyze how variations propagate through the attention layer. This makes it a valuable tool for understanding and enhancing Transformer robustness, particularly under adversarial settings.

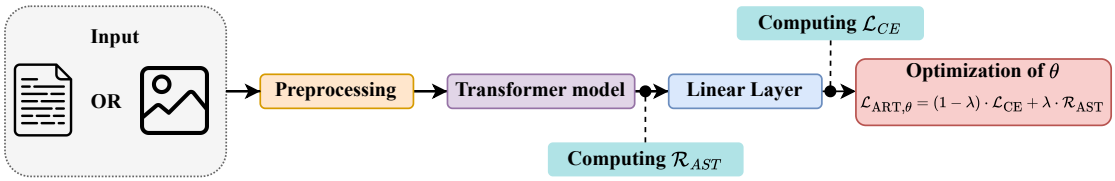

Figure 1: Overview of the ART training pipeline. Given text or image inputs, the data is first preprocessed and processed by a Transformer model followed by a linear classification. The cross-entropy loss $\mathcal{L}_{CE}$ is computed from the output logits, while the regularization term $\mathcal{R}_{AST}$ is computed from intermediate attention representations. The final loss $\mathcal{L}_{ART,\theta}$, combines both terms, weighted by a regularization parameter λ, and is used to optimize the model parameters θ.

4.1.3 Robustness Analysis with AST:

In the following, we establish formal robustness guarantees for Transformer self-attention mechanisms using the AST. A proof of this result can be found in Appendix A.

Proposition 1. Let $X \in \mathcal{X}$ be an input matrix and $\delta \in \mathbb{R}^{N\times d}$ a perturbation. Define the perturbed input as $X' = X + \delta$, where $X' \in \mathbb{R}^{N\times d}$. Let $\mathcal{Z}[i,f] : \mathcal{X} \to \mathbb{R}$ denote the (i,f)-th scalar component of the self-attention map defined in Eq. (1), with $(i,f) \in \{1,\dots,N\} \times \{1,\dots,d_v\}$.

The following bound holds:

$$\mathcal{Z}[i,f](X) - \sum_{j=1}^{N}\sum_{g=1}^{d} \mathcal{A}[i,f,j,g]\,|\delta[j,g]| \le \mathcal{Z}[i,f](X') \le \mathcal{Z}[i,f](X) + \sum_{j=1}^{N}\sum_{g=1}^{d} \mathcal{A}[i,f,j,g]\,|\delta[j,g]|$$

The AST provides certified, element-wise robustness guarantees by bounding the influence of each input feature on attention outputs. Minimizing AST values reduces local sensitivity, yielding more stable representations under perturbations. To complement this results, we conduct theoretical and empirical robustness evaluations using the certified radius obtained from the AST (see Appendix D.1)

4.2 Analytical Expression of AST

While the *AST* is defined abstractly in Eq. (2), we now derive its explicit form for the self-attention mechanism introduced in Eq. (1). A detailed proof of this result, as well as the algorithms used to compute the AST, can be found in the Appendix (C.1,C.2). Moreover, as discussed in Appendix C.3, the *AST* generalizes the Lipschitz constant by expressing sensitivity as a tensor, capturing local and direction-dependent variations across input dimensions for a finer measure of robustness.

Proposition 2. Let $\mathcal{Z} : \mathcal{X} \to \mathbb{R}^{N\times d_v}$ be the self-attention mapping defined in Eq. (1), and let $X \in \mathcal{X}$ be an input matrix. Then, each component $\mathcal{A}[i,f,j,g]$ of the *Attention Sensitivity Tensor* can be computed as:

$$\mathcal{A}[i,f,j,g] = \max\left(\left|\underline{\mathcal{A}}[i,f,j,g]\right|,\ \left|\overline{\mathcal{A}}[i,f,j,g]\right|\right), \tag{3}$$

where

$$\underline{\mathcal{A}}[i,f,j,g] = \min\left(\mathcal{J}[i,f,j,g](X) \mid X \in \mathcal{X}\right),$$

$$\overline{\mathcal{A}}[i,f,j,g] = \max\left(\mathcal{J}[i,f,j,g](X) \mid X \in \mathcal{X}\right),$$

and the map $X \mapsto \mathcal{J}[i,f,j,g](X)$ is defined as:

$$\mathcal{J}[i,f,j,g](X) = \Bigg|\sum_{m=1}^{N}\Big(\mathcal{C}[i,m](X)\mathbf{1}_{(j,m)}(W_V)[g,f] + \mathcal{C}[i,m](X)V[m,f](X)(\mathcal{B}_m - \sum_{l=1}^{N}\mathcal{C}[i,l](X)\mathcal{B}_l)\Big)\Bigg|$$

With:

- $\mathcal{C}(X) = \text{softmax}\left(\frac{XW_Q(XW_K)^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{N\times N}$, where $\mathcal{C}[i,m](X) \in [0,1]$ denotes the (i,m)-th entry of the attention matrix (similarly for $\mathcal{C}[i,l](X)$), corresponding to the attention weight assigned at position i to input position m.
- $(W_V)[g,f] \in \mathbb{R}$ is the (g,f)-th entry of the learned value projection matrix W_V.
- $V[m,f](X) \in \mathbb{R}$ denotes the (m,f)-th entry of the projected value matrix $V = XW_V \in \mathbb{R}^{N\times d_v}$.
- For $p \in \{m,l\}$, the scalar $\mathcal{B}_p \in \mathbb{R}$ is given by:

$$\mathcal{B}_p = \frac{1}{\sqrt{d_k}}\Big[\mathbf{1}_{(j,i)}\cdot((W_QW_K^\top)X[p,:]^\top)[g] + \mathbf{1}_{(j,p)}\cdot(X[i,:]W_QW_K^\top)[g]\Big]$$

where $X[p,:] \in \mathbb{R}^d$ denotes the p-th row of the input matrix, $((W_QW_K^\top)X[p,:]^\top)[g] \in \mathbb{R}$ and

$(X[i,:]W_Q W_K^\top)[g] \in \mathbb{R}$ denote the g-th components of the projected key–query at position p, and i, respectively.

where $i, j \in \{1, \ldots, N\}$, $f \in \{1, \ldots, d_v\}$, and $g \in \{1, \ldots, d\}$. The indicator function $\mathbf{1}_{\{a=b\}}$ is defined as 1 if $a = b$, and 0 otherwise.

4.3 Attention-Regularized Transformer (ART) Framework

Building on the explicit AST formulation, we propose the ***Attention-Regularized Transformer (ART)***, a training framework that incorporates the AST defined in Eq.(3) directly into the objective function as a regularization term. The overall training procedure is illustrated in Figure 1. The ART loss function is defined as:

$$\mathcal{L}_{\text{ART},\theta} = (1 - \lambda) \cdot \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{R}_{\text{AST}},$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, $\lambda > 0$ is a regularization weight, θ denotes model parameters, and $\mathcal{R}_{\text{AST}}$ is the AST-based regularization term, defined as:

$$\mathcal{R}_{\text{AST}} = \sum_{i=1}^{N} \sum_{f=1}^{d_v} \sum_{j=1}^{N} \sum_{g=1}^{d} \mathcal{A}[i, f, j, g]. \quad (4)$$

This regularization penalizes high sensitivity in the attention mechanism by minimizing the magnitudes of the AST entries. As a result, the model tends to learn smoother attention behaviors that are less sensitive to small perturbations in the input.

Incorporating AST into the training objective yields a principled and tractable framework for enhancing Transformer robustness without sacrificing expressiveness, applicable to both text and image domains through task-specific AST computation.

We study the effect of the regularization weight λ, which balances $\mathcal{L}_{\text{CE}}$ and $\mathcal{R}_{\text{AST}}$. As shown in Appendix D.2, ART maintains strong robustness and clean accuracy across a broad range of λ values.

Algorithm 1 summarizes the ART training process. The objective combines the cross-entropy loss $\mathcal{L}_{\text{CE}}$ with the AST-based regularizer $\mathcal{R}_{\text{AST}}$, weighted by λ. For each batch, predictions are obtained and $\mathcal{L}_{\text{CE}}$ is computed. For every self-attention layer, the local AST is calculated using Algorithm 2 (in Appendix C.2), and aggregated across layers. The final loss $\mathcal{L}_{\text{ART}}$ is minimized by updating model parameters θ via gradient descent.

Algorithm 1 ART Training Framework

Require: Transformer with L self-attention layers, training set $\mathcal{D}_{\text{train}}$, regularization weight λ
Ensure: Parameters θ minimizing the AST regularized loss

for each batch $(x, y) \in \mathcal{D}_{\text{train}}$ **do**
  $\hat{y} \leftarrow \text{model}(x)$; $\mathcal{L}_{\text{CE}} \leftarrow \text{CE}(\hat{y}, y)$
  $\mathcal{R}_{\text{AST}} \leftarrow 0$
  for self-attention layer $\ell = 1, \ldots, L$ **do**
    Compute $\mathcal{A}_\ell[i, f, j, g]$ using Algorithm 2 (in Appendix C.2)
    $\mathcal{R}_{\text{AST}} \mathrel{+}= \sum_{i,j,f,g} \mathcal{A}_\ell[i, f, j, g]$
  end for
  $\mathcal{L}_{\text{ART}} \leftarrow (1 - \lambda)\mathcal{L}_{\text{CE}} + \lambda \mathcal{R}_{\text{AST}}$
  Update model parameters θ
end for

5 Experiments

In this section, we evaluate ART on text and image classification tasks, comparing its robustness against multiple adversarial attacks and defense baselines, and analyzing its efficiency and impact on attention sensitivity. Appendix D provides detailed experimental settings and analyses, including certified robustness bounds from the AST (D.1), a sensitivity study of the regularization weight λ (D.2), layer-wise AST analysis showing reduced Jacobian bounds under ART (D.3), and dataset statistics with implementation details for reproducibility (D.5, D.6).

5.1 Experimental Setup

5.1.1 Datasets

To evaluate the multi-modal robustness of ART, we consider both text and image classification benchmarks. For text, we use *IMDB* (Maas et al., 2011), a binary sentiment dataset, and *QNLI* (Wang et al., 2018), a binary entailment task from GLUE derived from SQuAD. For image, we adopt *CIFAR-10/100* (Krizhevsky et al., 2009) 60,000 images across 10/100 classes and *Imagenette* (Howard, 2019), a 10-class ImageNet subset for efficient evaluation.

5.1.2 Models

For text, we adopt BERT-base model (Devlin et al., 2019), which has 12 Transformer encoder layers with 12 self-attention heads each, fine-tuned following standard text classification practice. For image classification, we evaluate ART using three Transformer-based architectures: the Vision Trans-

former (ViT-Small) (Dosovitskiy et al., 2020), the Data-efficient Image Transformer (DeiT-Tiny) (Touvron et al., 2021), and the Convolutional Vision Transformer (ConViT-Tiny) (d'Ascoli et al., 2021). All models are trained using the original default settings.

5.1.3 Evaluation Settings

We evaluate model robustness using adversarial attacks. For text, we apply three token-level attacks: *TextBugger* (Li et al., 2018), *TextFooler* (Jin et al., 2020), and *BERT-Attack* (Li et al., 2020), implemented via the TextAttack framework (Morris et al., 2020). To ensure a fair comparison with the baselines, and due to their high computational cost, adversarial examples are generated on 1,000 randomly selected test samples per dataset. The detailed attack configurations are reported in Appendix D. For image, we employ the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) and Projected Gradient Descent (PGD-2) (Madry et al., 2017), both constrained by an ℓ_∞ perturbation bound of $8/255$.

We assess model performance using two standard metrics: *Clean Accuracy*, measuring the proportion of correctly classified samples on the original test set, and *Robust Accuracy* (RA), which quantifies the model's ability to maintain correct predictions under adversarial perturbations:

$$\text{RA} = \frac{1}{|\mathcal{D}|} \sum_{(X,y)\in\mathcal{D}} \mathbf{1}\big[f(X+\delta(X)) = y\big],$$

where $\mathcal{D}$ is the test set, f the classifier, y the true label, and $\delta(X)$ an adversarial perturbation generated by a fixed attack method. The indicator function $\mathbf{1}[\cdot]$ returns 1 if the prediction remains correct after perturbation and 0 otherwise.

We also provide theoretical and empirical robustness analyses based on the certified radius derived from the AST (Appendix D.1).

5.1.4 Baseline Models

We compare ART against representative baselines from both text and image domains, grouped by widely used robustness techniques.

- **Adversarial Training Methods:** These methods use training on adversarially perturbed samples. This group includes PGD-K (Madry et al., 2017), FreeLB (Zhu et al., 2019), FreeLB++ (Li et al., 2021), and TA-VAT (Li and Qiu, 2020).

Method	IMDB			
	Standard	TextFooler	BERT-Attack	TextBugger
baseline (BERT)	92.1	10.3	5.8	5.3
MixADA (Si et al., 2020)	91.9	19.0	7.6	11.5
PGD-K (Madry et al., 2017)	93.2	26.0	21.0	18.9
FreeLB (Zhu et al., 2019)	93.0	29.0	21.7	22.9
TA-VAT (Li and Qiu, 2020)	93.0	28.0	19.2	22.8
InfoBERT (Wang et al., 2020)	92.0	29.2	30.7	25.4
DNE (Zhou et al., 2020)	90.4	28.0	27.0	26.5
ASCC (Dong et al., 2021)	87.8	19.4	11.0	14.1
SAFER (Ye et al., 2020)	**93.5**	39.5	38.5	40.0
RanMASK (Zeng et al., 2021)	93.2	22.0	36.9	18.0
FreeLB++ (Li et al., 2021)	93.2	45.3	40.6	36.6
Text-CRS (Zhang et al., 2024)	91.5	<u>84.4</u>	–	–
EarlyRobust (Xi et al., 2022)	91.8	49.7	43.8	46.8
RSMI (Moon et al., 2023)	92.2	56.4	51.1	<u>54.4</u>
DSRM (Gao et al., 2023)	<u>93.4</u>	56.3	54.1	67.2
SMAAT (Altinisik et al., 2024)	92.2	77.9	<u>60.8</u>	–
ART (Ours)	91.9	**85.3** (↑ 0.9)	**72.9** (↑ 12.1)	**78.3** (↑ 11.1)

Table 1: Performance (%) of *ART* with different defense methods on the IMDB dataset using BERT. The best results are in **bold**, second-best are <u>underlined</u>. The last row of each block indicates the gains of the RA between our method and the best baseline.

- **Adversarial Data Augmentation:** MixADA (Si et al., 2020) augments training with semantically and syntactically transformed inputs.
- **Smoothness and Certifiable Defenses:** InfoBERT (Wang et al., 2020), DSRM (Gao et al., 2023), RSMI (Moon et al., 2023), and SMAAT (Altinisik et al., 2024) use regularization, smoothing, or certification.
- **Consistency-Based and Contrastive Learning:** EarlyRobust (Xi et al., 2022), ASCC (Dong et al., 2021), and Text-CRS (Zhang et al., 2024) deal with perturbations through consistency training and contrastive learning objectives.
- **Adversarial Detection or Masking Strategies:** DNE (Zhou et al., 2020), SAFER (Ye et al., 2020), and RanMASK (Zeng et al., 2021) defend by identifying or masking perturbed tokens or features.
- **Regularized-based Models:** LipsFormer (Qi et al., 2023), L2Former (Kim et al., 2021), LN-Former (Dasoulas et al., 2021) and SpecFormer (Hu et al., 2024) enforce Lipschitz continuity bounds to provide robustness against adversarial perturbations.

5.2 Robustness on text classification tasks

We evaluate ART on text classification tasks under multiple adversarial attacks. Across all datasets, ART consistently improves robustness while maintaining or even improving clean accuracy.

Method	QNLI			
	Standard	TextFooler	BERT-Attack	TextBugger
baseline (BERT)	90.6	5.8	3.5	10.9
PGD-K (Madry et al., 2017)	90.6	14.3	17.3	27.9
FreeLB (Zhu et al., 2019)	90.7	12.8	21.4	29.8
InfoBERT (Wang et al., 2020)	90.4	18.0	13.1	15.4
FreeLB++ (Li et al., 2021)	**91.1**	16.4	20.7	30.2
DSRM (Gao et al., 2023)	90.1	27.6	20.4	37.1
ART (Ours)	90.7	**64.5** (↑ 36.9)	**43.4** (↑ 22.0)	**42.3** (↑ 5.2)

Table 2: Performance (%) of *ART* and prior defense methods on the QNLI dataset using BERT. The best results are in **bold**, second-best are underlined. Final row reports ART's improvement over the strongest baseline.

5.2.1 Main Results

As shown in Tables 1 and 2, ART outperforms all baselines on both IMDB and QNLI under multiple adversarial attacks. On IMDB, it achieves the highest accuracy under TextFooler (85.3%), BERT-Attack (72.9%), and TextBugger (78.3%), exceeding DSRM by 29.0%, 18.8%, and 11.1%, respectively. On QNLI, ART delivers the best robustness, improving by 36.9%, 22.0%, and 5.2% over prior methods under the same attacks. These results demonstrate that ART strengthens text robustness without compromising clean accuracy.

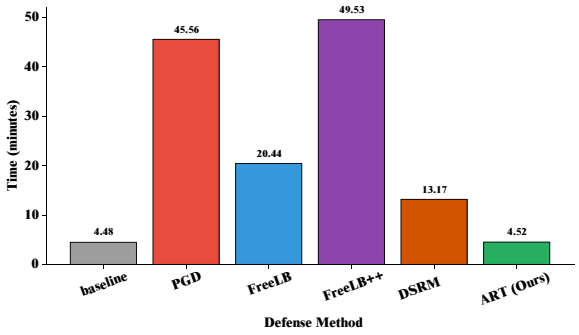

Figure 2: Training time per epoch (in minutes) for different defense methods on the IMDB dataset using BERT.

5.2.2 Training Efficiency

Beyond robustness, we assess the training efficiency of ART on IMDB with a BERT backbone. As shown in Figure 2, adversarial methods such as PGD and FreeLB++ require 45.56 and 49.53 minutes per epoch, respectively, whereas ART trains in 4.52 minutes comparable to standard fine-tuning. This yields over 10× and 11× speedups relative to FreeLB and FreeLB++, while attaining superior robustness, underscoring the practicality of ART when both robustness and efficiency are critical.

5.3 Robustness on Image Classification Tasks

We evaluate ART on image classification benchmarks using Transformer-based models. Across datasets and under adversarial settings, ART consistently improves robustness while maintaining clean accuracy.

5.3.1 Main Results

Table 3 reports ART's performance on CIFAR-10, CIFAR-100, and Imagenette with ViT-S, DeiT-Ti, and ConViT-Ti backbones. ART consistently yields the highest robust accuracy under FGSM and PGD-2. For example, on CIFAR-10 with ConViT-Ti it attains 53.79% (FGSM) and 49.57% (PGD-2), surpassing the next best method by $> 5\%$ and $> 20\%$, respectively. On Imagenette, ART reaches 88.80% (FGSM) and 66.00% (PGD-2) while preserving high clean accuracy (98.80%). Comparable gains are observed on CIFAR-100. Overall, ART provides strong adversarial defense without sacrificing clean performance.

5.3.2 Training Efficiency

We evaluate ART's efficiency by measuring the time for 10 training steps, normalized to a vanilla ViT (1.0). As shown in Table 4, ART introduces only a modest overhead ($1.25\times$), substantially lower than LipsFormer ($5.5\times$), LNFormer ($4.5\times$), and even SpecFormer ($1.5\times$). These results underscore ART's scalability and training efficiency.

5.3.3 Impact of ART on Attention Sensitivity

To examine the source of ART's robustness, we analyze the behavior of AST. Figure 3 compares a vanilla ViT with ART-regularized model on CIFAR-10, showing that ART markedly reduces sensitivity magnitudes across all layers, thereby stabilizing attention responses to perturbations. For instance, at layer 6 the total AST drops from 833.38 to 0.0408, a reduction of over four orders of magnitude. This demonstrates how ART directly limits the sensitivity that adversarial inputs can exploit. Full per-layer AST values are provided in Appendix D.3.

5.3.4 Attack-dependent robustness behavior

The differences in robustness observed across adversarial attacks are consistent with the theoretical foundations of ART. As established in Section 4, the AST regularizer explicitly constrains the first-order sensitivity of the self-attention mechanism by bounding the Jacobian of attention outputs with

Model	Method	CIFAR-10			CIFAR-100			Imagenette		
		Standard	FGSM	PGD-2	Standard	FGSM	PGD-2	Standard	FGSM	PGD-2
ViT-S	LipsFormer (Qi et al., 2023)	71.13	31.48	4.17	40.05	9.92	1.36	86.80	36.60	30.20
	L2Former (Kim et al., 2021)	79.65	39.98	13.39	53.20	15.35	5.92	92.80	58.00	37.80
	LNFormer (Dasoulas et al., 2021)	75.82	33.72	7.75	48.81	13.04	7.27	92.00	49.00	41.80
	TransFormer (Dosovitskiy et al., 2020)	87.09	45.56	22.35	63.52	19.82	7.01	94.20	72.60	48.00
	SpecFormer (Hu et al., 2024)	**88.52**	50.58	29.53	**69.78**	23.92	9.67	97.20	84.20	61.60
	ART (Ours)	85.13	**51.16** (↑0.58)	**48.28** (↑18.75)	62.22	**25.45** (↑1.53)	**26.26** (↑16.59)	**99.20**	**85.20** (↑1.00)	**62.00** (↑0.40)
DeiT-Ti	LipsFormer (Qi et al., 2023)	72.54	36.14	3.46	39.72	8.66	0.96	79.40	30.60	8.00
	L2Former (Kim et al., 2021)	78.09	36.64	5.05	49.02	12.63	2.56	82.40	51.00	6.00
	LNFormer (Dasoulas et al., 2021)	77.16	34.78	3.81	52.50	14.40	2.83	80.20	44.80	9.20
	TransFormer (Dosovitskiy et al., 2020)	86.40	46.10	14.46	62.79	19.89	2.35	90.00	64.20	13.00
	SpecFormer (Hu et al., 2024)	**87.42**	45.71	18.10	**64.14**	20.93	1.61	92.20	70.20	26.20
	ART (Ours)	82.26	**47.40** (↑1.30)	**42.02** (↑23.92)	56.76	**22.26** (↑1.33)	**21.45** (↑18.62)	**97.20**	**72.40** (↑2.20)	**43.60** (↑17.40)
ConViT-Ti	LipsFormer (Qi et al., 2023)	79.71	38.47	7.09	48.08	10.84	1.57	90.60	44.40	24.60
	L2Former (Kim et al., 2021)	81.33	40.17	12.76	49.86	13.86	2.03	93.40	62.20	30.80
	LNFormer (Dasoulas et al., 2021)	75.48	28.67	3.56	51.13	11.62	2.41	84.20	36.20	15.00
	TransFormer (Dosovitskiy et al., 2020)	87.78	48.89	20.26	64.68	22.94	4.96	93.20	68.80	40.80
	SpecFormer (Hu et al., 2024)	87.49	47.64	20.53	65.57	21.78	4.10	92.60	69.00	28.60
	ART (Ours)	**97.06**	**53.79** (↑4.90)	**49.57** (↑29.04)	**85.65**	**29.97** (↑7.03)	**14.20** (↑9.24)	**98.80**	**88.80** (↑19.80)	**66.00** (↑25.20)

Table 3: Robustness comparison of ART against baseline Transformer variants on CIFAR-10, CIFAR-100, and Imagenette under standard and adversarial settings. The best results are in **bold**. Reported gains in the final row of each block indicate the improvement of ART over the most competitive baseline.

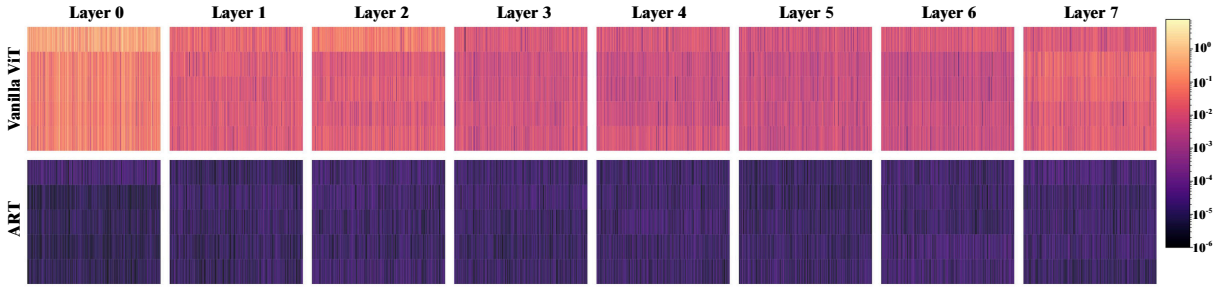

Figure 3: Comparison of AST across transformer layers for a vanilla ViT and our ART-based model on CIFAR-10. ART reduces attention sensitivity throughout all layers, as shown by the lower magnitudes (logarithmic scale).

Transformer	LipsFormer	L2Former	LNFormer	SpecFormer	ART
1.0	5.5	2.5	4.5	1.5	**1.25**

Table 4: Relative running time to complete 10 training steps under standard training. All values are normalized with respect to the vanilla Transformer (ViT) baseline.

respect to input perturbations. Consequently, attacks that primarily exploit local, first-order gradient information are expected to benefit most directly from AST-based regularization.

This behavior is clearly reflected in our results. In the image domain, FGSM is a single-step attack that follows the sign of the loss gradient, while in the text domain, TextFooler similarly relies on local gradient-based importance scores to guide token substitutions. Because both attacks operate within a locally linear approximation of the model, reducing attention sensitivity leads to substantial robustness gains under these settings.

In contrast, PGD is an iterative multi-step attack that progressively explores a broader region of the loss landscape beyond a single gradient step. Since ART primarily targets local sensitivity through first-order bounds, its impact on PGD robustness is naturally more moderate. Nevertheless, we consistently observe improved robustness under PGD across models and datasets, indicating that limiting local attention sensitivity still contributes positively even against stronger iterative attacks.

For text-based attacks such as BERT-Attack and TextBugger, robustness differences arise not from higher-order gradient effects but from the underlying perturbation mechanisms. These attacks generate adversarial examples through masked language model substitutions or combined lexical–semantic

transformations rather than direct gradient ascent in embedding space. As a result, each attack induces a distinct perturbation distribution, leading to varying robustness gains from AST regularization.

Finally, the magnitude of improvement also depends on dataset-specific sensitivity characteristics. For example, CIFAR-10 and CIFAR-100 differ substantially in inter-class similarity and patch-level variability, resulting in different baseline attention sensitivities and thus different absolute robustness gains when AST is minimized. A similar phenomenon is observed in text classification, where linguistic structure and label semantics shape the perturbation distributions of different datasets. These findings further support the interpretation that ART improves robustness by systematically reducing attention sensitivity in a manner that interacts with both attack design and dataset structure.

6 Discussion

The results show that AST-based regularization is improves Transformer's robustness. By penalizing attention sensitivity, ART promotes stable attention patterns and achieves consistently higher robust accuracy than strong baselines on both text and image tasks, while maintaining competitive clean accuracy. For example, on IMDB and QNLI, ART substantially improves robustness under attacks such as TextFooler, while on CIFAR-100 and Imagenette it outperforms Lipschitz regularized models, confirming effectiveness beyond text. Importantly, ART remains robust across a range of regularization weights λ: although extreme values slightly reduce clean accuracy, robust accuracy stays high (e.g., $> 82\%$ on IMDB and $> 60\%$ on QNLI in all cases; detailed analysis in Appendix D.2). Figure 3 further illustrates reduced attention instability across layers, explaining the reliability of predictions under adversarial perturbations. Compared to costly, task-specific adversarial training, ART is efficient, architecture-agnostic, and broadly applicable, providing a simple yet effective way to enhance robustness. These findings support that reducing attention sensitivity via AST is both theoretically sound and practically effective.

7 Conclusion

We presented *ART*, a unified and efficient framework for improving Transformer robustness in both text and image tasks. By minimizing the *Attention Sensitivity Tensor* (AST), ART stabilizes attention outputs under perturbations, yielding stronger robustness without compromising clean accuracy. Experiments across diverse benchmarks show that ART outperforms strong baselines while remaining efficient and compatible with existing architectures.

Limitations

While ART shows strong results, our current experiments focus only on classification tasks. Extending the method to generation (e.g., summarization (Thota and Nilizadeh, 2024)) or cross-modal tasks (Yin et al., 2023) remains an important direction. This work evaluates ART on widely used text and image benchmarks. Extending the analysis to larger models (Kumar and Mishra, 2025), or additional robustness settings (Wang and Zhao, 2024) is left for future work.

References

Enes Altinisik, Safa Messaoud, Husrev Taha Sencar, Hassan Sajjad, and Sanjay Chawla. 2024. Explaining the role of intrinsic dimensionality in adversarial training. *arXiv preprint arXiv:2405.17130*.

Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee.

George Dasoulas, Kevin Scaman, and Aladin Virmaux. 2021. Lipschitz normalization for self-attention layers with application to graph neural networks. In *International Conference on Machine Learning*, pages 2456–2466. PMLR.

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.

Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. 2021. Towards robustness against natural language word substitutions. *arXiv preprint arXiv:2107.13541*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. 2021. Convit: Improving vision transformers with

soft convolutional inductive biases. In *International conference on machine learning*, pages 2286–2296. PMLR.

SongYang Gao, Shihan Dou, Yan Liu, Xiao Wang, Qi Zhang, Zhongyu Wei, Jin Ma, and Ying Shan. 2023. Dsrm: Boost textual adversarial training with distribution shift risk minimization. *arXiv preprint arXiv:2306.15164*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Jeremy Howard. 2019. The imagenette dataset. *URL https://github. com/fastai/imagenette*.

Xixu Hu, Runkai Zheng, Jindong Wang, Cheuk Hang Leung, Qi Wu, and Xing Xie. 2024. Specformer: Guarding vision transformer robustness via maximum singular value penalization. In *European Conference on Computer Vision*, pages 345–362. Springer.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. 2021. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pages 5562–5571. PMLR.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Pankaj Kumar and Subhankar Mishra. 2025. Robustness in large language models: A survey of mitigation strategies and evaluation metrics. *arXiv preprint arXiv:2505.18658*.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*.

Linyang Li and Xipeng Qiu. 2020. Tavat: Token-aware virtual adversarial training for language understanding. *arXiv preprint arXiv:2004.14543*.

Zongyi Li, Jianhan Xu, Jiehang Zeng, Linyang Li, Xiaoqing Zheng, Qi Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021. Searching for an effective defender: Benchmarking defense against adversarial word substitution. *arXiv preprint arXiv:2108.12777*.

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 *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. 2022. Towards robust vision transformer. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 12042–12051.

Pierre-Jean Meyer, Alex Devonport, and Murat Arcak. 2021. *Interval reachability analysis: Bounding trajectories of uncertain systems with boxes for control and verification*. Springer Nature.

Han Cheol Moon, Shafiq Joty, Ruochen Zhao, Megh Thakkar, and Xu Chi. 2023. Randomized smoothing with masked inference for adversarially robust text classifications. *arXiv preprint arXiv:2305.06522*.

John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*.

Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, and Lei Zhang. 2023. Lipsformer: Introducing lipschitz continuity to vision transformers. *arXiv preprint arXiv:2304.09856*.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Better robustness by more coverage: Adversarial training with mixup augmentation for robust fine-tuning. *arXiv preprint arXiv:2012.15699*.

Poojitha Thota and Shirin Nilizadeh. 2024. Attacks against abstractive text summarization models through lead bias and influence functions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13727–13741, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2020. Infobert: Improving robustness of language models from an information theoretic perspective. *9th International Conference on Learning Representations (ICLR)*.

Yuqing Wang and Yun Zhao. 2024. Rupbench: Benchmarking reasoning under perturbations for robustness evaluation in large language models. *arXiv preprint arXiv:2406.11020*.

Zhiheng Xi, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Efficient adversarial training with robust early-bird tickets. *arXiv preprint arXiv:2211.07263*.

Mao Ye, Chengyue Gong, and Qiang Liu. 2020. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3465–3475, Online. Association for Computational Linguistics.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. 2023. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36:52936–52956.

Jiehang Zeng, Xiaoqing Zheng, Jianhan Xu, Linyang Li, Liping Yuan, and Xuanjing Huang. 2021. Certified robustness to text adversarial attacks by randomized [MASK]. *arXiv preprint arXiv:2105.03743*.

Xinyu Zhang, Hanbin Hong, Yuan Hong, Peng Huang, Binghui Wang, Zhongjie Ba, and Kui Ren. 2024. Text-crs: A generalized certified robustness framework against textual adversarial attacks. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 2920–2938. IEEE.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-wei Chang, and Xuanjing Huang. 2020. Defense against adversarial attacks in nlp via dirichlet neighborhood ensemble. *arXiv preprint arXiv:2006.11627*.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.

A Robustness Analysis with AST

This section offers the detailed proof of Proposition 1, showing how the Attention Sensitivity Tensor (AST) bounds the influence of input perturbations on attention outputs, thereby providing certified robustness guarantees.

A.1 Proof of Proposition 1

Fix an output coordinate $(i, f) \in \{1, \ldots, N\} \times \{1, \ldots, d_v\}$. Let $X \in \mathcal{X} \subseteq \mathbb{R}^{N \times d}$ be an input matrix and $\delta \in \mathbb{R}^{N \times d}$ a perturbation. Define the perturbed input as $X' = X + \delta$.

By the multivariate mean value theorem, there exists a point $C \in \mathcal{X}$ on the segment joining X and X' such that:

$$\mathcal{Z}[i, f](X') - \mathcal{Z}[i, f](X) = \left\langle \nabla \mathcal{Z}[i, f](C),\ X' - X \right\rangle,$$

where $\nabla \mathcal{Z}[i, f](C)$ denotes the gradient of $\mathcal{Z}[i, f]$ with respect to $X \in \mathbb{R}^{N \times d}$, evaluated at the intermediate point C, and $\langle \cdot, \cdot \rangle$ is the inner product over $\mathbb{R}^{N \times d}$.

Expanding the inner product, one gets,

$$\mathcal{Z}[i, f](X') - \mathcal{Z}[i, f](X) = \sum_{j=1}^{N} \sum_{g=1}^{d} \frac{\partial \mathcal{Z}[i, f](C)}{\partial X[j, g]} \cdot \delta[j, g].$$

Taking the absolute value and applying the triangle inequality:

$$\left| \mathcal{Z}[i, f](X') - \mathcal{Z}[i, f](X) \right| \leq \sum_{j=1}^{N} \sum_{g=1}^{d} \left| \frac{\partial \mathcal{Z}[i, f](C)}{\partial X[j, g]} \right| \cdot \left| \delta[j, g] \right|.$$

By the definition of the Attention Sensitivity Tensor (AST), we have:

$$\left| \frac{\partial \mathcal{Z}[i, f](C)}{\partial X[j, g]} \right| \leq \mathcal{A}[i, f, j, g], \quad \forall (i, f, j, g) \in \mathcal{I}, \quad \forall C \in \mathcal{X}.$$

where $\mathcal{I} = \{1, \ldots, N\} \times \{1, \ldots, d_v\} \times \{1, \ldots, N\} \times \{1, \ldots, d\}$.
Substituting this into the inequality above yields:

$$\left| \mathcal{Z}[i, f](X') - \mathcal{Z}[i, f](X) \right| \leq \sum_{j=1}^{N} \sum_{g=1}^{d} \mathcal{A}[i, f, j, g] \cdot \left| \delta[j, g] \right|.$$

Therefore, we obtain the certified bound for all $(i, f) \in \{1, \ldots, N\} \times \{1, \ldots, d_v\}$ and all $X \in \mathcal{X} \subseteq \mathbb{R}^{N \times d}$,

$$\begin{aligned} \mathcal{Z}[i, f](X) - \sum_{j=1}^{N} \sum_{g=1}^{d} \mathcal{A}[i, f, j, g] \cdot \left| \delta[j, g] \right| &\leq \mathcal{Z}[i, f](X') \\ &\leq \mathcal{Z}[i, f](X) + \sum_{j=1}^{N} \sum_{g=1}^{d} \mathcal{A}[i, f, j, g] \cdot \left| \delta[j, g] \right|. \end{aligned}$$

□

B Auxiliary Results for AST computation

This section provides auxiliary mathematical results required to bound the intermediate quantities involved in the computation of the Attention Sensitivity Tensor (AST) described in Eq. (3). These results form the technical basis for the algorithms presented later in Appendix C.2.

First, we present a result on how to compute the interval bounds for affine functions, which serves as a basis for bounding linear projections such as queries, keys, and values in the attention mechanism.

Proposition 3: Consider the map

$$\begin{aligned}\varphi : \mathcal{X} &\to \mathbb{R}, \\ x &\mapsto a^\top x,\end{aligned}$$

where $\mathcal{X} = \prod_{r=1}^{d} [\underline{x}_r, \overline{x}_r] \subset \mathbb{R}^d$ and $a \in \mathbb{R}^d$ is a fixed row vector. Then φ admits interval bounds

$$\varphi(x) \in [\, \underline{a^\top x}, \overline{a^\top x} \,], \quad \forall x \in \mathcal{X},$$

with

$$\underline{a^\top x} = \sum_{r=1}^{d} a_r \cdot \begin{cases} \underline{x}_r & \text{if } a_r \geq 0, \\ \overline{x}_r & \text{if } a_r < 0, \end{cases} \quad \overline{a^\top x} = \sum_{r=1}^{d} a_r \cdot \begin{cases} \overline{x}_r & \text{if } a_r \geq 0, \\ \underline{x}_r & \text{if } a_r < 0. \end{cases}$$

Next, we extend the previous result to the softmax operation by deriving explicit interval bounds on the attention coefficients.

Proposition 4: (Bounds on Softmax Coefficients Under Logit Intervals) Let $N \in \mathbb{N}$ be fixed. Consider a logit matrix $Z \in \mathbb{R}^{N \times N}$. For a fixed index $i \in \{1, \ldots, N\}$, denote by

$$Z_{i,:} = (Z_{i,1}, Z_{i,2}, \ldots, Z_{i,N}) \in \mathbb{R}^N$$

the i-th row of Z.

Let the *softmax map* acting on the i-th row be defined as

$$\begin{aligned} C : \mathbb{R}^N &\to \mathbb{R}^N, \\ Z[i,:] &\mapsto C[i,:], \end{aligned}$$

with entries

$$C[i,m](Z) := \frac{\exp(Z_{i,m})}{\sum_{\ell=1}^{N} \exp(Z_{i,\ell})}, \quad \forall m \in \{1, \ldots, N\}.$$

Assume that each logit component $Z_{i,m}$ is known only to lie within a closed interval

$$Z_{i,m} \in [\, \underline{Z}_{i,m},\ \overline{Z}_{i,m} \,], \quad \forall m \in \{1, \ldots, N\},$$

where $\underline{Z}_{i,m}, \overline{Z}_{i,m} \in \mathbb{R}$ satisfy $\underline{Z}_{i,m} \leq \overline{Z}_{i,m}$. Define the interval matrix as:

$$[\, \underline{Z}, \overline{Z} \,] := \{\, Z \in \mathbb{R}^{N \times N} \mid \underline{Z}_{i,m} \leq Z_{i,m} \leq \overline{Z}_{i,m},\ \forall i, m \,\}.$$

Then, for each $m \in \{1, \ldots, N\}$, the corresponding softmax coefficient $C[i,m](Z)$ satisfies

$$\frac{e^{\underline{Z}_{i,m}}}{\sum_{\ell=1}^{N} e^{\overline{Z}_{i,\ell}}} \ \leq\ C[i,m](Z)\ \leq\ \frac{e^{\overline{Z}_{i,m}}}{\sum_{\ell=1}^{N} e^{\underline{Z}_{i,\ell}}}, \quad \forall Z \in [\, \underline{Z}, \overline{Z} \,].$$

Equivalently, defining the lower and upper softmax bounds as

$$\underline{C}[i,m](\underline{Z}, \overline{Z}) := \frac{e^{\underline{Z}_{i,m}}}{\sum_{\ell=1}^{N} e^{\overline{Z}_{i,\ell}}}, \qquad \overline{C}[i,m](\underline{Z}, \overline{Z}) := \frac{e^{\overline{Z}_{i,m}}}{\sum_{\ell=1}^{N} e^{\underline{Z}_{i,\ell}}}.$$

Proof of Proposition 4. Let $Z \in \mathbb{R}^{N\times N}$ be a matrix with entries $Z_{i,m}$, and define the softmax coefficients as

$$C[i,m](Z) \;=\; \frac{\exp(Z_{i,m})}{\sum_{\ell=1}^{N}\exp(Z_{i,\ell})}, \quad \text{for all } i,m \in \{1,\dots,N\}.$$

Assume that Z is elementwise bounded by two matrices $\underline{Z}, \overline{Z} \in \mathbb{R}^{N\times N}$, that is,

$$Z \in [\,\underline{Z}, \overline{Z}\,] \quad\Longleftrightarrow\quad \underline{Z}_{i,m} \le Z_{i,m} \le \overline{Z}_{i,m} \quad \text{for all } i,m.$$

Fix $i \in \{1,\dots,N\}$ and $m \in \{1,\dots,N\}$.

First, we bound the numerator. Because the exponential function is strictly increasing on $\mathbb{R}$, from $Z_{i,m} \in [\underline{Z}_{i,m}, \overline{Z}_{i,m}]$ we obtain

$$e^{\underline{Z}_{i,m}} \;\le\; e^{Z_{i,m}} \;\le\; e^{\overline{Z}_{i,m}}.$$

Next, we bound the denominator. Monotonicity of the exponential and additivity yield

$$\sum_{\ell=1}^{N} e^{\underline{Z}_{i,\ell}} \;\le\; \sum_{\ell=1}^{N} e^{Z_{i,\ell}} \;\le\; \sum_{\ell=1}^{N} e^{\overline{Z}_{i,\ell}}.$$

We now pass to reciprocals. Since $x \mapsto 1/x$ is strictly decreasing on $\mathbb{R}_{>0}$, taking reciprocals reverses the inequalities and gives

$$\frac{1}{\sum_{\ell=1}^{N} e^{\overline{Z}_{i,\ell}}} \;\le\; \frac{1}{\sum_{\ell=1}^{N} e^{Z_{i,\ell}}} \;\le\; \frac{1}{\sum_{\ell=1}^{N} e^{\underline{Z}_{i,\ell}}}.$$

Finally, we combine the bounds. The softmax coefficient can be written as

$$C[i,m](Z) = \frac{e^{Z_{i,m}}}{\sum_{\ell=1}^{N} e^{Z_{i,\ell}}}.$$

Multiplying the lower bound on the numerator with the lower bound (in the sense of reciprocals) on the denominator, and similarly for the upper bounds, we obtain

$$\frac{e^{\underline{Z}_{i,m}}}{\sum_{\ell=1}^{N} e^{\overline{Z}_{i,\ell}}} \;\le\; C[i,m](Z) \;\le\; \frac{e^{\overline{Z}_{i,m}}}{\sum_{\ell=1}^{N} e^{\underline{Z}_{i,\ell}}}.$$

Therefore, for all $Z \in [\,\underline{Z}, \overline{Z}\,]$, each softmax coefficient satisfies

$$C[i,m](Z) \in \left[\frac{e^{\underline{Z}_{i,m}}}{\sum_{\ell=1}^{N} e^{\overline{Z}_{i,\ell}}},\; \frac{e^{\overline{Z}_{i,m}}}{\sum_{\ell=1}^{N} e^{\underline{Z}_{i,\ell}}}\right].$$

□

Finally, we recall the main operations of interval arithmetic to make the paper self-contained. The following results can be found for example in (Meyer et al., 2021).

Proposition 5 (Interval arithmetic toolkit): Let $[a,\overline{a}], [b,\overline{b}] \subset \mathbb{R}$ be closed intervals with $a \le \overline{a}$ and $b \le \overline{b}$.

(i) Sum and difference.

$$[a,\overline{a}] + [b,\overline{b}] \;=\; [\,a+b,\; \overline{a}+\overline{b}\,], \qquad [a,\overline{a}] - [b,\overline{b}] \;=\; [\,a-\overline{b},\; \overline{a}-b\,].$$

(ii) Product of two intervals.

$$[a,\overline{a}] \cdot [b,\overline{b}] \;=\; \big[\,\min\{ab,\; a\overline{b},\; \overline{a}b,\; \overline{a}\,\overline{b}\},\; \max\{ab,\; a\overline{b},\; \overline{a}b,\; \overline{a}\,\overline{b}\}\,\big].$$

(iii) Dot product bound. Let $u_t \in [\underline{u}_t, \overline{u}_t]$ and $v_t \in [\underline{v}_t, \overline{v}_t]$ for $t = 1, \ldots, d_k$. Then

$$\sum_{t=1}^{d_k} u_t v_t \;\in\; \sum_{t=1}^{d_k} \big[\, \min\{\underline{u}_t\underline{v}_t,\ \underline{u}_t\overline{v}_t,\ \overline{u}_t\underline{v}_t,\ \overline{u}_t\overline{v}_t\},\ \max\{\underline{u}_t\underline{v}_t,\ \underline{u}_t\overline{v}_t,\ \overline{u}_t\underline{v}_t,\ \overline{u}_t\overline{v}_t\} \,\big].$$

(iv) Product with an interval difference. If $x \in [\underline{x}, \overline{x}]$, $y \in [\underline{y}, \overline{y}]$, and $s \in [\underline{s}, \overline{s}]$, then

$$x\,(y - s) \;\in\; [\underline{x}, \overline{x}] \cdot [\,\underline{y} - \overline{s},\ \overline{y} - \underline{s}\,],$$

with the product computed via *(ii)*.

(v) Sum of interval products. If $c_\ell \in [\underline{c}_\ell, \overline{c}_\ell]$ and $b_\ell \in [\underline{b}_\ell, \overline{b}_\ell]$ for $\ell = 1, \ldots, N$, then

$$\sum_{\ell=1}^{N} c_\ell b_\ell \;\in\; \sum_{\ell=1}^{N} \big[\, \min\{\underline{c}_\ell\underline{b}_\ell,\ \underline{c}_\ell\overline{b}_\ell,\ \overline{c}_\ell\underline{b}_\ell,\ \overline{c}_\ell\overline{b}_\ell\},\ \max\{\underline{c}_\ell\underline{b}_\ell,\ \underline{c}_\ell\overline{b}_\ell,\ \overline{c}_\ell\underline{b}_\ell,\ \overline{c}_\ell\overline{b}_\ell\} \,\big].$$

(Each summand is bounded by (ii), then summed by (i).)

C Analytical Expression of AST

This section provides the full analytical derivation of the Attention Sensitivity Tensor (AST) introduced in Eq. (2). Building on the self-attention definition in Eq. (1), we derive a closed-form expression for each tensor component $\mathcal{A}[i, f, j, g]$ and discuss its theoretical relationship with Lipschitz constant. These results form the computational and theoretical foundation of the ART regularizer.

We first present the detailed proof of Proposition 2, which derives the explicit analytical form of the AST defined in Eq. (3).

C.1 Proof of Proposition 2 (AST)

Consider the self-attention map $\mathcal{Z} : \mathcal{X} \to \mathbb{R}^{N \times d_v}$, defined for all $i \in \{1, \ldots, N\}$ and for all $f \in \{1, \ldots, d_v\}$ by:

$$\mathcal{Z}[i, f](X) = \sum_{m=1}^{N} \mathcal{C}[i, m](X) \cdot V[m, f](X),$$

where the input matrix $X \in \mathcal{X} \subseteq \mathbb{R}^{N \times d}$, and $\mathcal{C}(X) = \mathrm{softmax}\left(\frac{Q(X)K^\top(X)}{\sqrt{d_k}}\right)$ denotes the attention matrix and $V(X) = XW_V \in \mathbb{R}^{d \times d_v}$ is the value projection.

For all $i, j \in \{1, \ldots, N\}$, $f \in \{1, \ldots, d_v\}$, and $g \in \{1, \ldots, d\}$, and by the product rule, we have the following:

$$\frac{\partial \mathcal{Z}[i, f]}{\partial X[j, g]} = \sum_{m=1}^{N} \frac{\partial \mathcal{C}[i, m]}{\partial X[j, g]} \cdot V[m, f](X) + \sum_{m=1}^{N} \mathcal{C}[i, m](X) \cdot \frac{\partial V[m, f]}{\partial X[j, g]}.$$

Now we compute each term in this expression.

First, since $V(X) = XW_V$, we have:

$$\frac{\partial V[m, f]}{\partial X[j, g]} = \mathbf{1}_{(j,m)} \cdot (W_V)[g, f].$$

Next, to compute $\dfrac{\partial \mathcal{C}[i, m]}{\partial X[j, g]}$, recall that the attention weights arise from a softmax over scaled dot products:

$$\mathcal{C}[i, m](X) = \mathrm{softmax}(Z_{i,m}) = \frac{\exp(Z_{i,m})}{\sum_{l=1}^{N} \exp(Z_{i,l})},$$

with

$$Z_{i,m} = \frac{1}{\sqrt{d_k}}(Q_i \cdot K_m) = \frac{1}{\sqrt{d_k}}(X[i, :]W_Q \cdot X[m, :]W_K^\top).$$

Differentiating $Z_{i,m}$ with respect to $X[j,g]$, we get:

$$\mathcal{B}_m = \frac{\partial Z_{i,m}}{\partial X[j,g]} = \frac{1}{\sqrt{d_k}} \left[\mathbf{1}_{(j,i)} \cdot ((W_Q W_K^\top) X[m,:]^\top)[g] + \mathbf{1}_{(j,m)} \cdot (X[i,:] W_Q W_K^\top)[g] \right].$$

Now, using the derivative of softmax, we obtain:

$$\frac{\partial \mathcal{C}[i,m]}{\partial X[j,g]} = \mathcal{C}[i,m](X) \cdot \left(\mathcal{B}_m - \sum_{l=1}^{N} \mathcal{C}[i,l](X) \cdot \mathcal{B}_l \right).$$

Putting it all together, the partial derivative becomes:

$$\frac{\partial \mathcal{Z}[i,f]}{\partial X[j,g]} = \sum_{m=1}^{N} \mathcal{C}[i,m](X) \cdot V[m,f](X) \cdot \left(\mathcal{B}_m - \sum_{l=1}^{N} \mathcal{C}[i,l](X) \cdot \mathcal{B}_l \right) + \sum_{m=1}^{N} \mathcal{C}[i,m](X) \cdot \mathbf{1}_{(j,m)} \cdot (W_V)[g,f].$$

Taking the absolute value of this expression yields the Jacobian magnitude at index (i,f,j,g):

$$\mathcal{J}[i,f,j,g](X) = \left| \sum_{m=1}^{N} \left[\mathcal{C}[i,m](X) V[m,f](X) \left(\mathcal{B}_m - \sum_{l=1}^{N} \mathcal{C}[i,l](X) \mathcal{B}_l \right) + \mathcal{C}[i,m](X) \mathbf{1}_{(j,m)} (W_V)[g,f] \right] \right|$$

Finally, since the AST must upper-bound the Jacobian over the entire input domain $\mathcal{X}$, we define:

$$\mathcal{A}[i,f,j,g] = \max \left(|\underline{\mathcal{A}}[i,f,j,g]|, \ |\overline{\mathcal{A}}[i,f,j,g]| \right),$$

where

$$\underline{\mathcal{A}}[i,f,j,g] = \min_{X \in \mathcal{X}} \mathcal{J}[i,f,j,g](X), \quad \overline{\mathcal{A}}[i,f,j,g] = \max_{X \in \mathcal{X}} \mathcal{J}[i,f,j,g](X).$$

This completes the construction of the AST and concludes the proof. □

The computation of the *Attention Sensitivity Tensor (AST)* for self-attention layers follows a clear and structured procedure based on analytical bounds and practical algorithms. For each Jacobian entry (i,f,j,g), the process starts by bounding the linear projections of the input set $\mathcal{X}$ using *Proposition 3*, which gives intervals for $Q = XW_Q$, $K = XW_K$, and $V = XW_V$. These intervals are used to compute the scaled dot-product logits $Z_{i,m}$ through *Algorithm 3* and *Proposition 5(iii)*. Then, the attention weights $\mathcal{C}[i,m]$ are bounded using *Algorithm 4* and *Proposition 4*.

Next, the weighted average $S = \sum_\ell \mathcal{C}[i,\ell] B_\ell$ is computed with *Algorithm 5* using standard interval arithmetic from *Proposition 5*. Each term $T_m = \mathcal{C}[i,m]\, V[m,f]\, (B_m - S)$ is then bounded with *Algorithm 6*, combining the intervals of $\mathcal{C}[i,m]$, $V[m,f]$, and $(B_m - S)$. Finally, the total Jacobian interval $[\underline{\mathcal{J}}[i,f,j,g], \overline{\mathcal{J}}[i,f,j,g]]$ is obtained using *Algorithm 7*, which merges the effects of both the *context path* (through $\mathcal{C}[i,m]$ and $V[m,f]$) and the *value path* (through W_V).

The overall procedure is summarized in **Algorithm 2**, which loops over all layers and component pairs (i,f,j,g) to compute the final sensitivity entries:

$$\mathcal{A}[i,f,j,g] = \max \left(|\underline{\mathcal{J}}[i,f,j,g]|, \ |\overline{\mathcal{J}}[i,f,j,g]| \right),$$

as defined in *Proposition 2*. Each element of the tensor is thus described by upper and lower bounds $\overline{\mathcal{A}}[i,f,j,g]$ and $\underline{\mathcal{A}}[i,f,j,g]$, which together form the complete Attention Sensitivity Tensor $\mathcal{A}$.

C.2 Algorithms

This subsection details the complete algorithms used to compute the Attention Sensitivity Tensor (AST). These procedures implement the theoretical formulations derived in Appendix B, enabling the practical calculation of interval bounds and tensor entries during training.

Next, we establish the relationship between the AST and the Lipschitz constant of the self-attention mapping. Specifically, we show that the global Lipschitz bound corresponds to the maximum value of the AST components $\mathcal{A}[i,f,j,g]$.

Algorithm 2 Computation of the Attention Sensitivity Tensor (AST)

Require: Input set $\mathcal{X} \in \mathbb{R}^{N \times d}$, model parameters (W_Q, W_K, W_V)
Ensure: Attention Sensitivity Tensor $\mathcal{A}[i, f, j, g]$

for each layer $\ell = 1, \ldots, L$ **do**
  for each pair (i, f) in $\{1, \ldots, N\} \times \{1, \ldots, d_v\}$ **do**
    for each (j, g) in $\{1, \ldots, N\} \times \{1, \ldots, d\}$ **do**
      Compute $[\underline{Z}_{i,m}, \overline{Z}_{i,m}]$ via Algorithm 3.
      Compute attention weight intervals $[\underline{C}_{i,m}, \overline{C}_{i,m}]$ via Algorithm 4.
      Compute weighted average $[\underline{S}, \overline{S}]$ via Algorithm 5.
      Compute each summand $[\underline{T}_m, \overline{T}_m]$ via Algorithm 6.
      Compute Jacobian interval $[\underline{J}[i, f, j, g], \overline{J}[i, f, j, g]]$ via Algorithm 7.
      Define AST entry $\mathcal{A}[i, f, j, g] = \max(|\underline{J}[i, f, j, g]|, |\overline{J}[i, f, j, g]|)$.
    end for
  end for
end for
return $\mathcal{A}$

Algorithm 3 Bounding the logits $Z_{i,m}$

Require: Input set $\mathcal{X} \in \mathcal{X} \subseteq \mathbb{R}^{N \times d}$, indices (i, m), matrices W_Q, W_K
Ensure: Interval $[\underline{Z}_{i,m}, \overline{Z}_{i,m}]$

Compute $Q_i = X[i, :]W_Q$, $K_m = X[m, :]W_K$.
For each coordinate t, bound $Q_i[t]$ and $K_m[t]$ using Proposition 3.
Apply Proposition 5(iii) to get $\langle Q_i, K_m \rangle$ interval.
Scale by $1/\sqrt{d_k}$ to obtain $[\underline{Z}_{i,m}, \overline{Z}_{i,m}]$.

Algorithm 4 Bounding the attention weights $\mathcal{C}[i, m]$

Require: Intervals $[\underline{Z}_{i,m}, \overline{Z}_{i,m}]$ for all $m = 1, \ldots, N$
Ensure: Interval $[\underline{C}_{i,m}, \overline{C}_{i,m}]$

Compute $\exp(\underline{Z}_{i,m})$ and $\exp(\overline{Z}_{i,m})$.
Bound the denominator: $\sum_\ell \exp(Z_{i,\ell}) \in [\sum_\ell \exp(\underline{Z}_{i,\ell}), \sum_\ell \exp(\overline{Z}_{i,\ell})]$.
Apply Proposition 4 to obtain

$$\mathcal{C}[i, m] \in \left[\frac{e^{\underline{Z}_{i,m}}}{\sum_\ell e^{\overline{Z}_{i,\ell}}}, \frac{e^{\overline{Z}_{i,m}}}{\sum_\ell e^{\underline{Z}_{i,\ell}}} \right].$$

Algorithm 5 Bounding the weighted average S

Require: Intervals $\mathcal{C}[i, \ell] \in [\underline{C}_{i,\ell}, \overline{C}_{i,\ell}]$ and $B_\ell \in [\underline{B}_\ell, \overline{B}_\ell]$
Ensure: Interval $[\underline{S}, \overline{S}]$

For each ℓ, compute interval product $\mathcal{C}[i, \ell]B_\ell$ using Proposition 5(ii).
Sum all intervals with Proposition 5(v) to obtain S.

Algorithm 6 Bounding each summand T_m

Require: Intervals B_m, S, $V[m, f]$, and $\mathcal{C}[i, m]$
Ensure: Interval $[\underline{T}_m, \overline{T}_m]$

Compute $B_m - S$ using Proposition 5(i).
Multiply by $V[m, f]$ using Proposition 5(iv).
Multiply by $\mathcal{C}[i, m]$ using Proposition 5(ii).

Algorithm 7 Bounding the Jacobian entry $J[i, f, j, g]$

Require: Intervals from Algorithms 3–6
Ensure: $[\underline{J}[i, f, j, g], \overline{J}[i, f, j, g]]$ and AST entry $A[i, f, j, g]$
1: Compute all T_m intervals via Algorithm 6.
2: Bound value-path term $\mathcal{C}[i, m]\delta_{j=m}(W_V)[g, f]$ using Proposition 5(ii).
3: Sum all intervals to obtain $J[i, f, j, g](X) \in [\underline{J}[i, f, j, g], \overline{J}[i, f, j, g]]$.

C.3 Relationship Between AST and the Lipschitz Constant

In the context of robustness analysis, the Lipschitz constant is widely used as a global measure of a model's sensitivity to input perturbations. It provides a uniform upper bound on the change in the output norm relative to the input norm, independent of direction. While effective in theory, this scalar bound is often loose and overly conservative in practice, especially for models like Transformers where sensitivity may vary significantly across different input dimensions or positions.

The Attention Sensitivity Tensor (AST), introduced in ART, offers a finer-grained alternative. Rather than summarizing sensitivity with a single scalar, the AST captures localized, coordinate-wise upper bounds on the AST of the self-attention map. This structure-aware formulation allows ART to characterize and regularize directional sensitivity, leading to tighter certified bounds and better interpretability.

We now formalize the link between the AST and the Lipschitz constant of the self-attention mapping, as stated in the following proposition.

C.3.1 Proposition 6:

Consider the self-attention mapping $\mathcal{Z} : \mathcal{X} \to \mathbb{R}^{N\times d_v}$ and let a tensor $\mathcal{A} \in \mathbb{R}^{N\times d_v\times N\times d}$ be its Attention Sensitivity Tensor (AST). Then

$$L = \max_{i,f} \max_{j,g} \mathcal{A}[i, f, j, g]$$

is a Lipschitz constant of the map $\mathcal{Z}$.

C.3.2 Proof of the Proposition 6:

Consider the map $\mathcal{Z} : \mathcal{X} \to \mathbb{R}^{N\times d_v}$, where $\mathcal{X} \subseteq \mathbb{R}^{N\times d}$. Each component $\mathcal{Z}[i, f] : \mathcal{X} \to \mathbb{R}$ is a scalar function. Given two inputs $X, X' \in \mathcal{X}$, define the perturbation $\delta = X' - X \in \mathbb{R}^{N\times d}$.

Under the infinity norm, we have:

$$\|\mathcal{Z}(X') - \mathcal{Z}(X)\|_\infty = \max_{i,f} \left|\mathcal{Z}[i, f](X') - \mathcal{Z}[i, f](X)\right|.$$

Now for each (i, f), the mean value theorem guarantees that:

$$\left|\mathcal{Z}[i, f](X') - \mathcal{Z}[i, f](X)\right| = \left|\sum_{j=1}^{N}\sum_{g=1}^{d} \frac{\partial \mathcal{Z}[i, f](C^{i,f})}{\partial X[j, g]} \cdot \delta[j, g]\right|,$$

for some intermediate point $C^{i,f} \in \mathcal{X}$ on the segment between X and X'.

Taking absolute values and applying the triangle inequality, one gets,

$$\begin{aligned}
\left|\mathcal{Z}[i, f](X') - \mathcal{Z}[i, f](X)\right| &\le \sum_{j=1}^{N}\sum_{g=1}^{d} \left|\frac{\partial \mathcal{Z}[i, f](C^{i,f})}{\partial X[j, g]}\right| \cdot |\delta[j, g]| \\
&\le \sum_{j=1}^{N}\sum_{g=1}^{d} \mathcal{A}[i, f, j, g] \cdot |\delta[j, g]| \\
&\le \max_{j,g} \mathcal{A}[i, f, j, g] \cdot \|\delta\|_\infty.
\end{aligned}$$

Taking the maximum over all (i, f), we obtain:

$$\|\mathcal{Z}(X') - \mathcal{Z}(X)\|_\infty \leq L \cdot \|X' - X\|_\infty,$$

where $L = \max_{i,f} \max_{j,g} \mathcal{A}[i, f, j, g]$. Hence, L is a Lipschitz constant of $\mathcal{Z}$. □

D Experiment Details

This section provides detailed experimental results and settings supporting our main findings. We begin with a certified robustness analysis based on our theoretical framework, followed by a hyper-parameter study to assess sensitivity. We then examine the stability benefits introduced by ART, report dataset statistics, and conclude with implementation and training details. Together, these sections offer a comprehensive view of our evaluation methodology and reproducibility setup.

D.1 Certified Robustness Analysis

We evaluate the certified robustness of ART using the theoretical framework introduced in Section 4. Specifically, we compute sample-wise robustness guarantees using the Attention Sensitivity Tensor (AST), which upper-bounds how much attention outputs can change with respect to perturbations in the input. This enables us to derive a certified ℓ_p-norm radius under which the model's prediction is guaranteed to remain stable. We first define the following Proposition 7:

D.1.1 Proposition 7: Certified Radius from AST.

Let $\mathcal{Z} : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times d_v}$ be the self-attention mapping as defined in Eq. (1), and let $\mathcal{A} \in \mathbb{R}^{N\times d_v\times N\times d}$ be the Attention Sensitivity Tensor (AST), such that for all $i, j \in \{1, \ldots, N\}$, $f \in \{1, \ldots, d_v\}$, and $g \in \{1, \ldots, d\}$, we have:

$$\left|\frac{\partial \mathcal{Z}[i, f]}{\partial X[j, g]}\right| \leq \mathcal{A}[i, f, j, g].$$

Let $h : \mathbb{R}^{N\times d_v} \to \mathbb{R}^C$ be the classifier that maps attention outputs to logits, and let $c = \arg\max_k h_k(X)$ be the predicted class. Define the classification margin as:

$$m(X) = h_c(X) - \max_{k\neq c} h_k(X).$$

Then, for any perturbation $\delta \in \mathbb{R}^{N\times d}$ satisfying $\|\delta\|_p \leq \epsilon$, the prediction remains unchanged, i.e., $h(X + \delta) = h(X)$, provided that:

$$\epsilon \leq \frac{m(X)}{\|\nabla_{\mathcal{Z}} h_c(X)^\top \cdot \mathcal{A}\|_q},$$

where $\nabla_{\mathcal{Z}} h_c(X) \in \mathbb{R}^{N\times d_v}$ is the Jacobian of the predicted class logit with respect to the attention output, and q is the dual norm of p such that $\frac{1}{p} + \frac{1}{q} = 1$.

D.1.2 Proof of Proposition 7.

From Proposition 1, each output coordinate $\mathcal{Z}[i, f]$ of the self-attention map satisfies:

$$|\mathcal{Z}[i, f](X + \delta) - \mathcal{Z}[i, f](X)| \leq \sum_{j=1}^{N} \sum_{g=1}^{d} \mathcal{A}[i, f, j, g] \cdot |\delta[j, g]|. \tag{5}$$

Let $h : \mathbb{R}^{N\times d_v} \to \mathbb{R}^C$ be the classifier mapping the attention outputs $\mathcal{Z}(X)$ to logits $h(X)$. h_c depends on X only through the intermediate variable $\mathcal{Z}(X)$. By the multivariate mean value theorem applied to the map h_c along the segment joining $\mathcal{Z}(X)$ and $\mathcal{Z}(X + \delta)$, there exists a point $C \in \mathcal{X}$, such that

$$h_c(X + \delta) - h_c(X) = \langle \nabla_{\mathcal{Z}} h_c(C),\ X + \delta - X \rangle,$$

where $\nabla_{\mathcal{Z}} h_c(C) \in \mathbb{R}^{N\times d_v}$ denotes the gradient of h_c with respect to its $\mathcal{Z}$-argument, evaluated at the intermediate point C, and $\langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbb{R}^{N\times d_v}$.

Expanding this inner product gives

$$h_c(X+\delta) - h_c(X) = \sum_{i=1}^{N}\sum_{f=1}^{d_v} \frac{\partial h_c}{\partial \mathcal{Z}[i,f]}(C)\ \big(\mathcal{Z}[i,f](X+\delta) - \mathcal{Z}[i,f](X)\big).$$

Taking absolute values and applying the triangle inequality yields

$$|h_c(X+\delta) - h_c(X)| \le \sum_{i=1}^{N}\sum_{f=1}^{d_v} \left|\frac{\partial h_c}{\partial \mathcal{Z}[i,f]}(C)\right| \cdot |\mathcal{Z}[i,f](X+\delta) - \mathcal{Z}[i,f](X)|\,. \tag{6}$$

Substituting the bound from Eq. (5) into Eq. (6) gives

$$|h_c(X+\delta) - h_c(X)| \le \sum_{i=1}^{N}\sum_{f=1}^{d_v} \left|\frac{\partial h_c}{\partial \mathcal{Z}[i,f]}(C)\right| \left(\sum_{j=1}^{N}\sum_{g=1}^{d} \mathcal{A}[i,f,j,g]\,|\delta[j,g]|\right).$$

Rearranging the sums to collect terms depending on $\delta[j,g]$, we obtain

$$|h_c(X+\delta) - h_c(X)| \le \sum_{j=1}^{N}\sum_{g=1}^{d} \left(\sum_{i=1}^{N}\sum_{f=1}^{d_v} \left|\frac{\partial h_c}{\partial \mathcal{Z}[i,f]}(C)\right| \mathcal{A}[i,f,j,g]\right)|\delta[j,g]|.$$

Defining the tensor contraction

$$(\nabla_{\mathcal{Z}} h_c(C)^\top \cdot \mathcal{A})[j,g] := \sum_{i=1}^{N}\sum_{f=1}^{d_v} \frac{\partial h_c}{\partial \mathcal{Z}[i,f]}(C)\,\mathcal{A}[i,f,j,g],$$

we can write the above bound compactly as

$$|h_c(X+\delta) - h_c(X)| \le \sum_{j=1}^{N}\sum_{g=1}^{d} \left|(\nabla_{\mathcal{Z}} h_c(C)^\top \cdot \mathcal{A})[j,g]\right| |\delta[j,g]|.$$

Applying Hölder's inequality with conjugate norms (p,q), such that $\frac{1}{p}+\frac{1}{q}=1$, gives

$$|h_c(X+\delta) - h_c(X)| \le \left\|\nabla_{\mathcal{Z}} h_c(C)^\top \cdot \mathcal{A}\right\|_q \cdot \|\delta\|_p.$$

To ensure that the prediction remains unchanged, we require this change to be smaller than the classification margin:

$$|h_c(X+\delta) - h_c(X)| < m(X).$$

Therefore, if

$$\epsilon \le \frac{m(X)}{\|\nabla_{\mathcal{Z}} h_c(C)^\top \cdot \mathcal{A}\|_q},$$

then for any perturbation δ with $\|\delta\|_p \le \epsilon$, the predicted logit h_c remains strictly greater than all other logits h_k, and the classifier output is preserved:

$$h(X+\delta) = h(X).$$

□

D.1.3 Empirical Certification Results.

For each test input, we compute the predicted class logit margin and the sensitivity term from the corollary above, approximated using the same gradient-based structure regularized during training. This yields a sample-wise certified radius ϵ in embedding space.

Certified accuracy is reported as the percentage of test samples whose predictions are provably invariant under perturbations of magnitude at most ϵ. As shown in Figure 4, ART achieves strong certified robustness on both IMDB and QNLI. On IMDB, the model achieves 99.4% certified accuracy at $\epsilon = 0.1$, 96.9% at $\epsilon = 0.5$, and 93.8% at $\epsilon = 1.0$. On QNLI, the model retains 98.8% certified accuracy at $\epsilon = 0.1$ and 89.4% at $\epsilon = 1.0$.

These results demonstrate that ART provides substantial certified robustness against continuous perturbations in embedding space, far exceeding the granularity and interpretability of discrete synonym-based certification approaches.

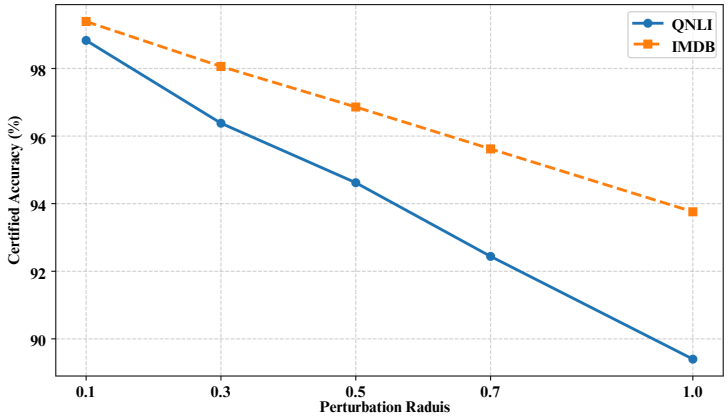

Figure 4: Certified accuracy of ART under ℓ_2 perturbations on IMDB and QNLI. Values reflect the percentage of test inputs for which predictions are provably invariant up to the specified certified radius ϵ.

D.1.4 Certified Neighborhoods of Synonyms

To provide a more intuitive interpretation of certified robustness, we visualize certified neighborhoods in the embedding space of individual words. Specifically, we consider synonym substitutions for the word *fantastic*, identified using Counter-fitted embeddings and filtered for contextual plausibility with BERT's masked language model. Figure 5 shows the projection of these synonyms into a two-dimensional space, with concentric circles denoting certified radii $\epsilon \in \{0.3, 0.5, 0.7, 1.0\}$. Synonyms lying inside a given radius correspond to substitutions that are guaranteed not to alter the model's prediction. This visualization highlights how the certified robustness framework translates into meaningful semantic neighborhoods in natural language.

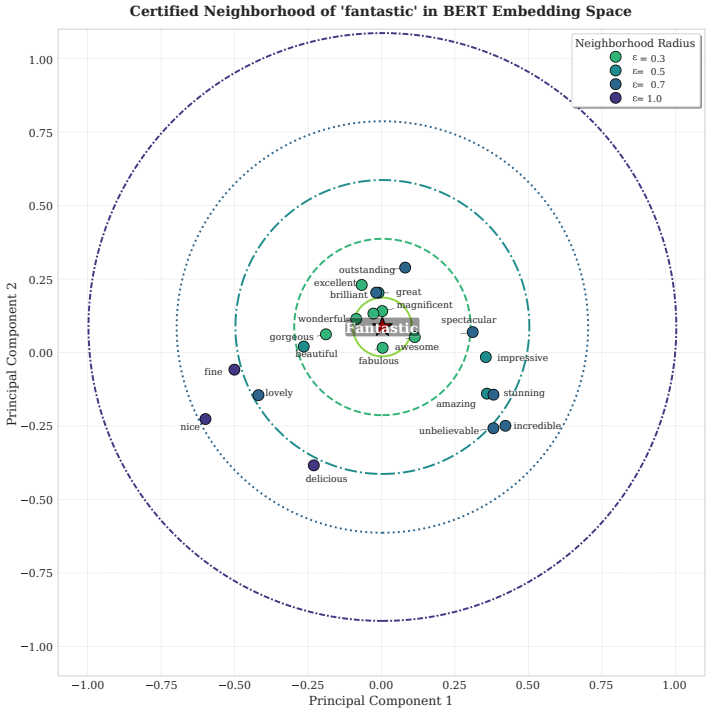

Figure 5: Certified synonym neighborhood of the word *fantastic*. Concentric circles correspond to certified radii ϵ. Synonyms within a circle are provably invariant substitutions at that robustness level.

Next, we study the effect of the regularization weight λ, which balances the cross-entropy loss and the AST-based regularizer. This analysis demonstrates the stability of ART across a wide range of λ values.

D.2 Hyper-parameter Study.

D.2.1 Text classification tasks

We study the effect of the hyperparameter λ in the ART loss, which balances the standard cross-entropy loss and the AST-based regularization. Figure 6 presents both clean accuracy and robust accuracy (RA) across a range of λ values on IMDB and QNLI under the TextFooler attack.

On IMDB, we observe that extreme values of λ (very low or very high) lead to drops in clean accuracy, while intermediate values yield a more favorable trade-off. Notably, values in the range $\lambda \in [0.1, 0.3]$ consistently achieve high clean accuracy (up to 99.72%) and robust accuracy above 82%. The best result is obtained at $\lambda = 0.6$, with 99.76% clean accuracy and 85.20% RA, demonstrating that AST regularization can enhance robustness without harming clean performance.

On QNLI, clean accuracy remains relatively stable around 90.5% across all λ, but RA is more sensitive to λ. The highest RA of 64.50% is reached at $\lambda = 0.8$, while still maintaining 90.66% clean accuracy. Even at less optimal λ values (e.g., 0.3 or 0.6), RA remains above 60%, confirming the robustness benefit of the ART regularization.

Overall, these results confirm that λ has a measurable impact on the robustness–accuracy trade-off, and that ART achieves strong performance across a broad range of settings, with RA remaining above 82% on IMDB and above 60% on QNLI for all tested λ values.

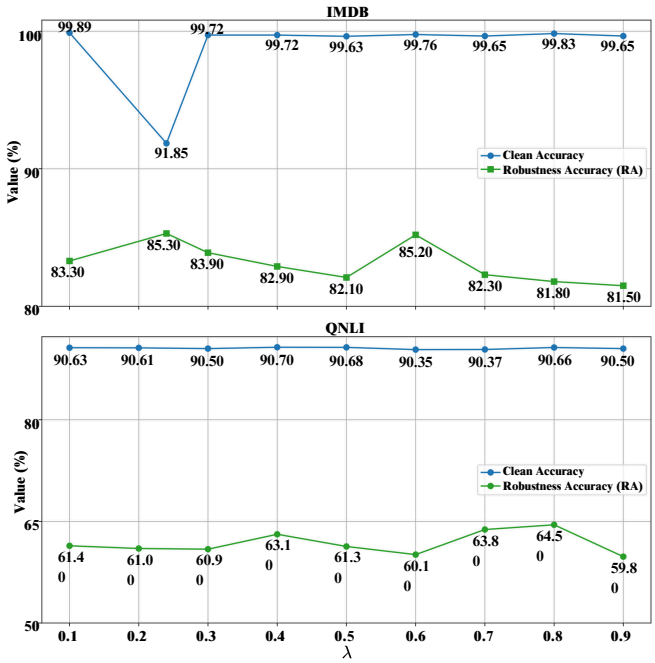

Figure 6: Effect of the trade-off parameter λ on clean accuracy and robust accuracy (RA) for IMDB and QNLI under TextFooler attack. Optimal robustness is achieved at moderate λ values.

Finally, we perform a layer-wise analysis of AST magnitudes to empirically verify that ART effectively reduces attention sensitivity throughout the Transformer architecture. These results quantitatively support the theoretical claim that minimizing $\mathcal{A}$ tightens the Jacobian bounds and stabilizes attention responses.

D.2.2 Image classification tasks

We further examine the influence of the trade-off parameter λ in the ART loss in the context of image classification. Table 5 summarizes standard (clean) accuracy as well as robust accuracy under FGSM and PGD-2 attacks for different values of λ.

λ	Standard	FGSM	PGD-2
0.1	83.49	48.08	46.17
0.2	83.71	48.36	46.72
0.3	84.32	49.29	46.88
0.4	84.19	49.79	46.78
0.5	84.62	48.67	47.45
0.6	85.13	**51.16**	**48.28**
0.7	84.92	49.35	47.40
0.8	**85.29**	49.66	47.21
0.9	85.15	48.26	46.92

Table 5: Effect of the trade-off parameter λ on standard accuracy and robust accuracy under FGSM and PGD-2 attacks for image classification.

Across the evaluated range, standard accuracy remains largely unaffected by the choice of λ, varying only slightly between 83.5% and 85.3%. This stability reflects the fact that incorporating AST-based regularization does not compromise clean performance, even when the regularization strength is increased.

Robust accuracy, however, shows a more pronounced sensitivity to λ. For FGSM, robustness improves as λ increases, reaching its highest value at $\lambda = 0.8$, after which a minor degradation is observed. A comparable pattern emerges under the PGD-2 attack, where robust accuracy increases up to $\lambda = 0.6$ and then gradually decreases for larger values.

Taken together, these observations indicate that intermediate values of λ provide a favorable balance between robustness and accuracy for image classification models. In particular, setting λ between 0.5 and 0.8 yields consistently improved robustness under both FGSM and PGD-2 while preserving competitive clean accuracy. The most favorable trade-off is observed around $\lambda = 0.6$–0.8, where robustness gains are maximized without noticeable loss in standard performance.

Overall, the results demonstrate that the choice of λ plays an important role in shaping adversarial robustness in image classification. Although overly large values may lead to diminishing returns, ART delivers reliable robustness improvements across a wide range of λ, with robust accuracy under both attacks remaining above 48% for appropriately selected settings.

D.3 Impact of ART on Attention Sensitivity

Table 6 provides a layer-wise comparison of the Attention Sensitivity Tensor (AST) values between the standard Vision Transformer (ViT) and the Attention-Regularized Transformer (ART). The vanilla ViT exhibits high AST values, particularly in the early layers (e.g., Layer 0), suggesting strong sensitivity of the attention mechanism to input perturbations. In contrast, the ART-regularized model demonstrates a consistent and substantial reduction in AST across all layers, indicating improved stability and robustness. This shows that ART helps the model learn more robust and consistent internal representations.

Layer	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Without ART	5381.81	865.51	998.99	990.60	1024.79	1026.09	833.38	824.38
With ART	0.0249	0.0283	0.0371	0.0324	0.0358	0.0293	0.0408	0.0366

Table 6: Sum of the Attention Sensitivity Matrix (AST) per layer, comparing vanilla ViT (without ART) and ART-regularized models. The significantly lower AST values under ART indicate reduced attention sensitivity.

D.4 Dataset Statistics

Table 7 summarizes the datasets used in our experiments across both text and image domains. For text, we use IMDB and QNLI, which are binary classification tasks with large-scale training sets. For image classification, we include CIFAR-10, CIFAR-100, and ImageNette, covering a range of dataset sizes and number of classes. This diverse set of benchmarks allows us to evaluate the effectiveness and generality of our method across different domains and levels of classification difficulty.

Domain	Dataset	Training set	Test set	Classes
Text	IMDB	25,000	25,000	2
	QNLI	104,743	5,463	2
Image	CIFAR-10	50,000	10,000	10
	CIFAR-100	50,000	10,000	100
	ImageNette	9,469	3,925	10

Table 7: Statistics of the datasets used in our experiments.

D.5 Detailed Setup

All experiments were conducted on a single NVIDIA A100 80GB GPU. Table 8 summarizes the ART training configuration for the BERT model, and for ViT, DeiT, and ConViT models.

D.6 Detailed Attack Configuration

The following table 9 summarizes the adversarial attack configurations used in our experiments. For each attack we report the substitution source, the maximum number of candidate replacements considered per token (top-k), the maximum fraction of tokens allowed to be changed in a sample, the tokenizer used

Domain	Model	Param.	IMDB	QNLI	CIFAR-10 / CIFAR-100	ImageNette
Text	BERT	Optimizer	Adam		–	–
		Batch size	64	16	–	–
		Hidden size	768		–	–
		Learning rate	$2e^{-5}$		–	–
		Max length	256		–	–
		Early stopping	Yes		–	–
Image	ViT-S / DeiT-Ti / ConViT-Ti	Optimizer	–	–	SGD	
		Batch size	–	–	64	
		Weight decay	–	–	$1e^{-5}$	
		Learning rate	–	–	$1e^{-1}$	
		Epochs	–	–	40	

Table 8: ART Training Setup for BERT on *IMDB* and *QNLI*, and for ViT-S, DeiT-Ti, and ConViT-Ti on *CIFAR-10*, *CIFAR-100*, and *ImageNette*.

(matched to the victim model), and brief implementation details. These settings were chosen to balance attack strength and semantic/fluency preservation; stopword protection and semantic filtering are applied where noted.

Attack	Substitution source	Max candidates / top-k	Max % tokens changed	Tokenizer	Configuration details
TextFooler	Counter-fitted embeddings	50	40%	victim tokenizer	Word-level synonym substitutions via nearest-neighbors in counter-fitted embedding space; stopwords protected.
BERT-Attack	BERT masked-language model	50	40%	victim tokenizer	MLM-based candidate generation using BERT; proposals ranked and filtered by recipe constraints to preserve fluency and semantics.
TextBugger	GloVe embeddings + character edits	8	40%	victim tokenizer	Hybrid character-level (typos) and embedding-based word swaps; semantic-similarity filtering and stopword protection enabled.

Table 9: Text-Attack configuration.

D.7 Computational Efficiency Analysis

Beyond wall-clock training time, we report system-level efficiency metrics for ART and baselines under identical hardware. To evaluate the computational footprint of ART across vision benchmarks, we extend our analysis to the CIFAR-10 dataset. All models were trained and profiled under identical hardware and batch configurations to ensure fair comparison.

Model	Params (M)	Inference FLOPs (G)	Training FLOPs (G)	Memory (GB)	Throughput (samples/s)
ViT-Small-ART	21.67	6.44	19.32	1.36	342.5
DeiT-Tiny-ART	5.53	1.83	5.48	1.47	991.6
ConViT-Tiny-ART	5.52	2.17	6.50	1.81	541.0

Table 10: System-level computational efficiency on CIFAR-10. Parameter counts, FLOPs, memory usage, and throughput are reported for ART-enhanced transformer models under identical hardware. ART maintains competitive efficiency while scaling effectively across model variants.

E Adversarial attack examples

This section illustrates representative adversarial examples for both text and image classification tasks. These visualizations demonstrate how imperceptible perturbations, such as synonym substitutions or pixel-level noise, can alter model predictions, highlighting the necessity of robustness-oriented approaches like ART.

We first present examples of adversarial perturbations in text, showing how minor synonym replacements can change the model's prediction while preserving semantic meaning.

E.1 Text classification tasks

Figure 7 illustrates representative adversarial examples in text domain. A synonym substitution attack is shown, where subtle word replacements preserve the original semantics but cause the model to flip its sentiment prediction.

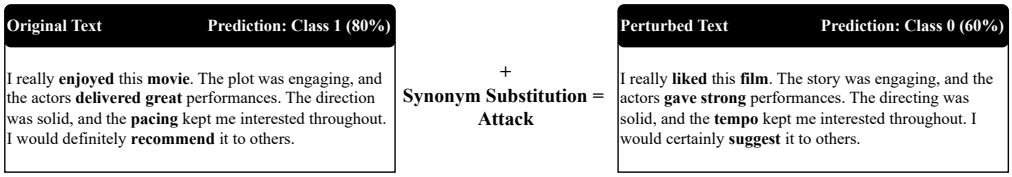

Figure 7: Example of a Synonym Substitution Attack on a Sentiment Classifier. The original review is slightly modified by replacing a few words with their synonyms (e.g., "enjoyed" → "liked", "movie" → "film"). Although the meaning stays the same, the model's prediction changes from positive (Class 1, 80%) to negative (Class 0, 60%). This shows how vulnerable NLP models can be to small, semantic-preserving adversarial attacks.

We then illustrate adversarial examples in the image domain, where subtle pixel-level perturbations cause incorrect classifications despite visually indistinguishable images.

E.2 Image classification tasks

Figure 8 illustrates representative adversarial examples in image domain. Imperceptible pixel-level perturbations are applied, leading to incorrect classifications while the visual content remains unchanged to the human eye.

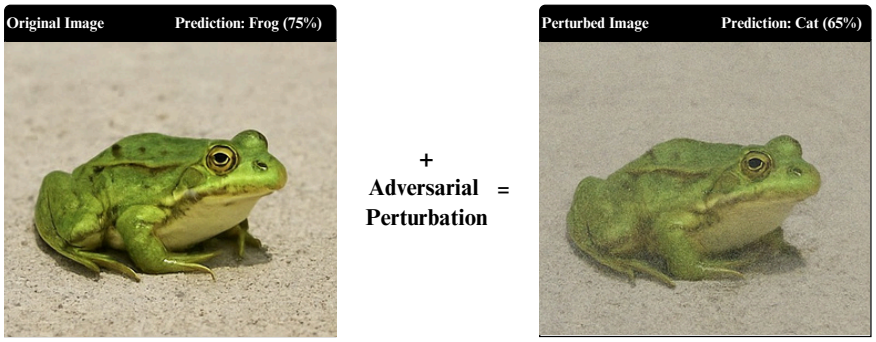

Figure 8: Example of an Adversarial Perturbation on a CIFAR-10 Image. A small, imperceptible perturbation is added to a clean image of a frog (predicted as Frog, 75%) to create an adversarial example. Although the perturbed image appears visually identical, the classifier's prediction flips to Cat (65%). This demonstrates the vulnerability of image classifiers to adversarial attacks that preserve human-perceived semantics.