ESCUELA DE
INGENIERÍA INFORMÁTICA

PONTIFICIA
UNIVERSIDAD
CATÓLICA DE
VALPARAÍSO

D-TRADES ENTRENAMIENTO ADAPTATIVO PARA UN BALANCE ÓPTIMO DE ROBUSTEZ ADVERSARIAL Y PRECISIÓN APLICADO A MODELOS DE VISIÓN POR COMPUTADORA

JORGE VILLARREAL GONZÁLEZ
ADEMIR MUÑOZ RODRÍGUEZ

PROFESOR GUÍA: EMANUEL VEGA

PROFESOR CORREFERENTE: MARCELO BECERRA

INFORME DE AVANCE
INGENIERÍA CIVIL INFORMÁTICA
NOVIEMBRE 2025

RESUMEN

El problema de trade-off en el contexto de ataques adversariales surge al aplicar métodos de defensa en redes neuronales convolucionales para aumentar su robustez adversarial. Si bien estos métodos fortalecen la resistencia del modelo, suelen provocar una disminución en la precisión bajo condiciones estándar, generando un desequilibrio que limita su aplicabilidad en escenarios reales. La mayoría de las investigaciones recientes se centran en incrementar la robustez adversarial, sin considerar adecuadamente el impacto en la precisión. En este trabajo se propone D-TRADES, un método orientado a optimizar el equilibrio entre precisión estándar y robustez adversarial, basado en el método TRADES, donde el parámetro de regularización λ se redefine como una función dinámica dependiente de la entrada. De esta forma, el modelo puede ajustar de manera adaptativa el peso entre muestras adversariales y estándar, con el objetivo de alcanzar un balance más eficiente entre robustez y precisión. Por medio de un estudio comparativo de métodos de defensa y trade-off asociado, con el fin de analizar sus efectos en el rendimiento general del modelo. Los resultados muestran que D-TRADES logra un equilibrio robustez-precisión, mantiene competitividad frente a métodos de defensa actuales y evalúa su desempeño frente ataques y métricas recientes de la literatura.

Palabras clave: trade-off, robustez, precision, adversarial

ABSTRACT

The trade-off problem in the context of adversarial attacks arises when applying defense methods in convolutional neural networks to increase their adversarial robustness. While these methods strengthen the model's resistance, they often cause a decrease in accuracy under standard conditions, creating an imbalance that limits their applicability in real-world scenarios. Most recent research focuses on increasing adversarial robustness without adequately considering the impact on accuracy. This paper proposes D-TRADES, a method aimed at optimizing the balance between standard accuracy and adversarial robustness, based on the TRADES method, where the regularization parameter λ is redefined as a dynamic function dependent on the input. In this way, the model can adaptively adjust the weight between adversarial and standard samples, with the aim of achieving a more efficient balance between robustness and accuracy. Through a comparative study of defense methods and associated trade-offs, we analyze their effects on the overall performance of the model. The results show that D-TRADES achieves a robustness-accuracy balance, remains competitive with current defense methods, and evaluates its performance against recent attacks and metrics in the literature.

Keywords: trade-off, robustness, accuracy, adversarial

ÍNDICE GENERAL

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

Los métodos de defensa han probado ser técnicas efectivas para mejorar la robustez adversarial de los modelos de aprendizaje profundo (DL), en contraste, estas conducen a una reducción indeseable de la precisión estándar (Cohen et al., 2019; Madry et al., 2018; Papernot et al., 2016; Y. Song et al., 2018; Xu et al., 2017), este fenómeno es el "equilibrio" o trade-off entre ambos objetivos.

La precisión estándar corresponde al rendimiento de un modelo con muestras limpias, es decir, muestras sin manipulaciones realizadas por terceros. En contraste, la robustez adversarial es la capacidad de un modelo para mantener predicciones correctas incluso cuando la muestra presenta pequeñas e imperceptibles perturbaciones diseñadas específicamente para engañar al modelo, conocidas como ataques adversariales (Rade and Moosavi-Dezfooli, 2021). La importancia en la búsqueda de la estabilidad entre estos objetivos radica en el apartado de la seguridad. Los ataques adversariales representan una amenaza real para la seguridad de los modelos de aprendizaje profundo desplegados en aplicaciones críticas. Ahora bien, como se mencionó anteriormente, si se robustece un modelo, provoca una reducción en la precisión estándar, es decir, si queremos que nuestro modelo funcione en un ámbito real, es crucial tener un modelo equilibrado (Arani et al., 2020). Por lo tanto, es relevante encontrar un método de defensa lo suficientemente eficiente para obtener un equilibrio entre la precisión estándar y la robustez adversarial.

Según Rade and Moosavi-Dezfooli, 2021 la defensa que ha demostrado ser la más efectiva es la del entrenamiento adversarial (AT), pero este modelo de defensa no se escapa respecto al problema del equilibrio. Debido a esto H. Zhang et al., 2019 propone TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization (TRADES), es un método de entrenamiento adversarial que busca equilibrar la precisión estándar y robustez adversarial mediante una función de perdida con dos términos, este equilibrio se controla con el hiperparámetro λ. Sin embargo, TRADES presenta dificultades, el rendimiento depende en parte de λ, si este no está bien balanceado, el modelo puede sobre ajustarse a las perturbaciones específicas usadas en el entrenamiento. Además, λ es particularmente sensible en datasets más complejos (ejemplo: CIFAR-100), y en ciertos casos puede inducir inestabilidad y sobreestimación de la robustez adversarial debido a gradient masking (Li et al., 2024).

A partir de estas limitaciones, esta investigación propone un método de defensa adversarial basado en TRADES, reemplazando el hiperparámetro λ fijo por un parámetro $\lambda(x)$ dinámico y dependiente de las características de cada muestra. Para evaluar esta propuesta, se realiza un análisis comparativo frente a métodos ampliamente utilizados TRADES, MART y FAAL empleando las arquitecturas ResNet-18 y VGG-16 sobre los datasets CIFAR-10 y MNIST. La metodología considera un análisis del estado del arte, un diseño experimental controlado y un conjunto de métricas que permiten evaluar rigurosamente la efectividad de la defensa propuesta en escenarios adversariales. En este sentido, el presente estudio no parte del supuesto de que la versión dinámica de $\lambda(x)$ supere a las defensas existentes, sino que busca evaluar su comportamiento real frente a diversas configuraciones adversariales y comprender en qué medida una regularización adaptativa modifica el trade-off característico de TRADES.

Finalmente, este documento se organiza en nueve capítulos. En primer lugar, la introducción presenta un contexto del problema, brechas en investigaciones previas y la metodología general a seguir. A continuación, los objetivos establecen los objetivos generales y secundarios de la investigación. Seguidamente, el marco teórico desarrolla los conceptos importantes que permiten entender el problema con un mayor grado de detalle. Más adelante, el estado del arte examina investigaciones similares, con énfasis en métodos de defensa y trade-off. En la quinta sección, la propuesta de solución explica con detalle el método de defensa propuesto, destacando diferencias con el enfoque base TRADES y presentando fórmulas y parámetros relevantes. Acto seguido, los experimentos muestran métricas comparativas entre la propuesta y métodos alternativos, comprobando efectividad y competitividad. Posteriormente, las conclusiones ofrecen una reflexión crítica sobre la propuesta a partir de los resultados obtenidos, así como posibles mejoras para futuras investigaciones. Finalmente la planificación muestra el cronograma seguido para completar la investigación por completo.

2. OBJETIVOS

En esta sección se detallan los objetivos del trabajo de investigación focalizado en el problema de trade-off. A continuación, se detalla el objetivo general, junto con los objetivos específicos.

2.1. Objetivo general

Proponer un método de trade-off basado en TRADES que utilice un parámetro de regularización dinámico dependiente de las muestras, con el fin de optimizar el equilibrio entre precisión estándar y robustez adversarial en redes neuronales convolucionales. El trabajo busca comparar la efectividad de este enfoque con otras técnicas de defensa existentes mediante un análisis experimental y el uso de métricas actuales de evaluación.

2.2. Objetivos específicos

Los objetivos de este trabajo de investigación son los siguientes:

- Analizar y sistematizar el estado del arte sobre redes neuronales convolucionales robustecidas frente a ataques adversariales, identificando y comprendiendo los principales métodos de trade-off, tipos de ataques, arquitecturas utilizadas, datasets de referencia y enfoques existentes.
- Entrenar redes neuronales convolucionales preentrenadas, sobre dos datasets distintas. Para establecer un punto de comparación sólido, se consideran como líneas base cuatro métodos de entrenamiento adversarial ampliamente reconocidos en la literatura, además de un modelo sin defensa alguna y la propuesta D-TRADES.
- Evaluar y comparar el rendimiento del método propuesto contra los otros modelos utilizando métricas concretas para la evaluación de los métodos de defensa. Permitiendo determinar si la propuesta equilibra eficazmente la precisión estándar y la robustez adversarial a la vez que se compara su rendimiento con los otros métodos de defensa.

3. MARCO TEÓRICO

El presente marco teórico conforma un conjunto de conceptos, definiciones y ejemplos para el entendimiento de la investigación. Inicialmente, se detallan conceptos generales sobre el marco de área de estudio, como lo serían el deep learning y las redes neuronales convolucionales. Posteriormente, se abarca la definición de conceptos específicos como muestras, ataques y defensas adversariales. Finalizando breve explicación del método TRADES, trade-off y de datasets a utilizar.

3.1. Deep Learning

El Deep Learning (DL) o Aprendizaje Profundo es un subcampo del Machine Learning (ML) que destaca por su relevancia en la extracción y análisis de datos. Este enfoque de la inteligencia artificial permite procesar grandes volúmenes de información (big data) de manera eficiente, resultando altamente efectivo en la identificación de patrones ocultos y en el entrenamiento de modelos complejos (Gheisari et al., 2023).

3.2. Redes Neuronales Convolucionales

Las redes neuronales convolucionales (CNN) son un tipo de redes de aprendizaje profundo, específicamente de visión por computadora. Diseñada para procesar datos con una estructura en forma de malla, como imágenes o señales bidimensionales. Y que, a diferencia de las redes neuronales tradicionales, estas incorporan capas convolucionales que aplican filtros (kernels) capaces de extraer automáticamente características tales como bordes, texturas y patrones (J. Song et al., 2019). Siendo además capaces de resolver efectivamente problemas complejos no lineales como la rotación y traslación de las imágenes. Esto se puede observar en la Figura 3.1, la estructura y funcionamiento de una CNN.

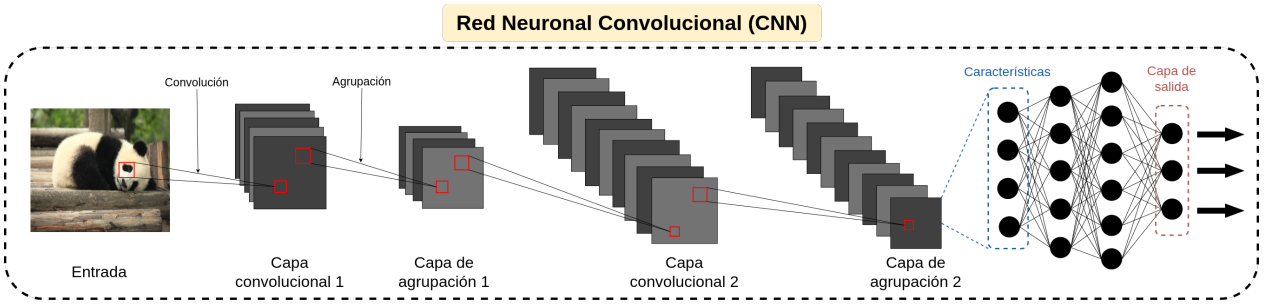

Figura 3.1: Estructura de una CNN.

3.2.1. Tipos de modelos CNN

Esta investigación considera el uso de diferentes arquitecturas de CNNs preentrenadas, entendidas como redes previamente entrenadas sobre grandes datasets, cuyos pesos pueden ser reutilizados

para resolver nuevas tareas, aprovechando el conocimiento previamente adquirido. De acuerdo con J. Song et al., 2019, es posible emplear una red preentrenada como extractor de características, utilizando dichas características para entrenar un clasificador independiente, o bien realizar un ajuste fino (fine-tuning) de la red para adaptarla a tareas específicas. Para la completitud de esta investigación, se emplearán dos de las arquitecturas mencionadas en Hassan et al., 2022.

- **Network-In-Network (NIN)**: Introduce un perceptrón multicapa en cada capa convolucional y reemplaza las capas totalmente conectadas por global average pooling, reduciendo el sobreajuste.
- **VGG-Net**: Aumenta la profundidad de la red hasta 19 capas con kernels 3×3, logrando mejoras de precisión y simplicidad estructural.
- **ResNet**: Utiliza conexiones residuales para facilitar el entrenamiento de redes muy profundas (>1,000 capas), resolviendo el problema de gradiente desvanecido.
- **DenseNet**: Introduce conexiones densas entre capas, reutilizando características, acelerando la convergencia y mejorando el flujo de gradientes.

3.3. Ataques Adversariales

Un ataque adversarial consiste en modificar una entrada legítima x añadiendo una perturbación pequeña e imperceptible para generar un ejemplo adversarial x' que induce a un modelo a predecir una etiqueta incorrecta. El objetivo del atacante es mantener la imagen visualmente similar al original, pero lo suficientemente alterada para cruzar la frontera de decisión del modelo. La siguiente ecuación representa la fórmula genérica de la definición de ataque adversarial (Dhamija and Bansal, 2024):

$$x' = x + \delta x \mid f(x') = y' \wedge f(x') \neq f(x) \tag{3.1}$$

La perturbación δx es un vector que indica cuánto se alteran los píxeles respecto a la imagen original. Su magnitud determina cuán lejos queda el ejemplo adversarial de su versión limpia. El tamaño permitido de la perturbación se controla mediante una restricción de norma: $||x-x'||_p \mid p \in \{0, 1, 2, \infty\}$; el límite más común es L_∞, que acota el cambio máximo permitido en cualquier píxel (Dhamija and Bansal, 2024). En la Figura 3.2 se observa un ejemplo respecto al ataque adversarial.

3.3.1. Tipos de Ataques Adversariales

Existiendo una gran cantidad de ellos, los cuales se diferencian en las siguientes categorías (Dhamija and Bansal, 2024):

- **Evasión**: Ocurre durante la fase de inferencia o despliegue, donde el atacante modifica las muestras de entrada para evadir al sistema, por medio de cambios imperceptibles para el ojo humano.
- **Extracción de modelos**: Son ataques de tipo caja negra, donde el atacante genera múltiples solicitudes/consultas con el fin de obtener información del modelo y con ello generar uno sustituto, que luego puede ser usado para generar un ataque más efectivo.
- **Envenenamiento de datos**: Se producen durante la fase de entrenamiento; el atacante puede inyectar datos venenosos al sistema con el fin de afectar la integridad del modelo.

- **Ataques dirigidos y no dirigidos**: La diferencia entre ambos es la siguiente.
  - **Ataques dirigidos**: Focalizado en inducir al modelo a clasificar una muestra adversarial dada por el atacante.
  - **Ataques no dirigidos**: Busca únicamente causar un error en la clasificación en el modelo.

3.3.2. Ataques Adversariales según nivel de conocimiento

Ademas de ello se dividen según el nivel de conocimiento del atacante (Ren et al., 2020), siendo ellos:

- **Caja Negra**: En un ataque de caja negra, el atacante solo puede realizar consultas al modelo objetivo, esto usualmente con la finalidad de hacer un modelo gemelo. Ejemplos de caja negra son: Ensemble Surrogate o Query-based attacks.
- **Caja Blanca**: En un ataque de caja blanca, el atacante tiene completamente conocimiento sobre el modelo objetivo, desde su arquitectura hasta sus parámetros, ejemplos de caja blanca son: Projected Gradient Descent (PGD), AutoAttack y FGSM.
- **Caja Gris**: Un ataque de caja gris, el atacante tiene conocimiento unicamente de la estructura.

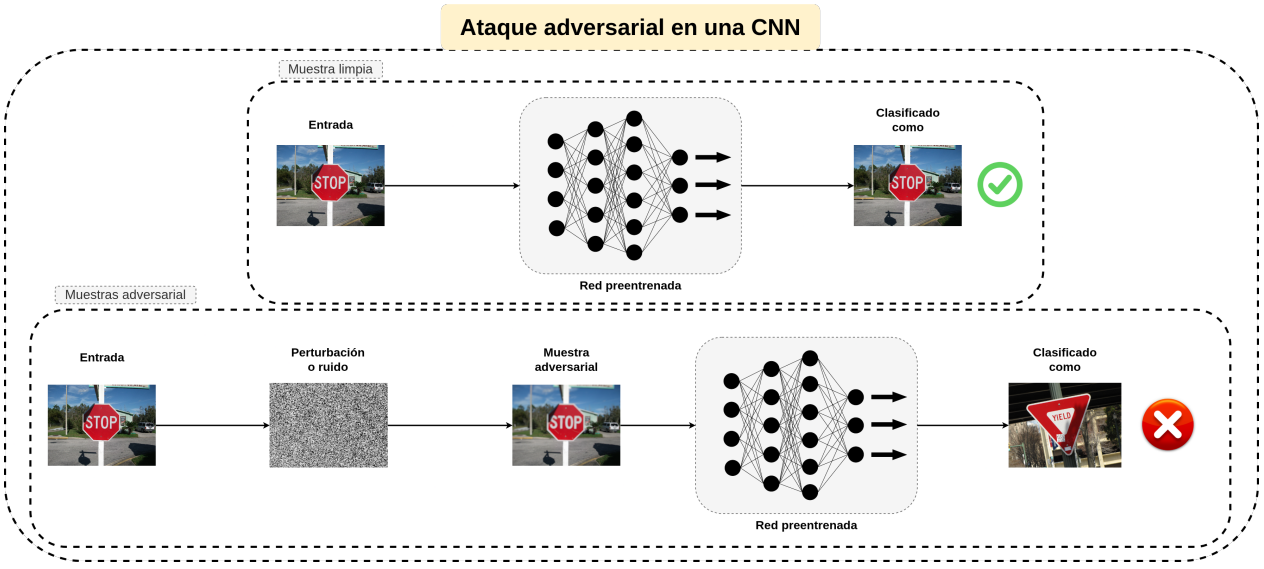

Figura 3.2: Clasificación de una CNN preentrenada frente a una entrada con y sin ruido.

3.4. Defensa Adversarial

Una defensa adversarial se define como una técnica o estrategia que permite a un modelo de aprendizaje profundo mantener su desempeño ante la presencia de muestras adversariales, reduciendo la efectividad de los ataques y aumentando su robustez adversarial. En la actualidad existen gran cantidad de métodos de defensas según (Wu et al., 2023), los cuales se separan en diferentes fases del ciclo de vida de una CNN.

- Pre-entrenamiento
- Entrenamiento
- Post-entrenamiento
- Despliegue
- Inferencia

3.4.1. Robustez Adversarial

Se define como la capacidad de un modelo de aprendizaje para mantener un desempeño confiable frente a ejemplos adversariales. Minimizando la pérdida adversarial esperada, considerando el peor caso de perturbaciones de un conjunto fijo, comúnmente perturbaciones acotadas como rotaciones, traslaciones o deformaciones espaciales suaves (Tsipras et al., 2019).

3.4.2. Fenómeno "trade-off" en visión por computadora

Este fenómeno ocurre al aplicar defensas robustas, es decir, técnicas que aumentan significativamente la resistencia de un modelo frente a ejemplos adversariales. Aunque estas defensas mejoran la robustez adversarial, suelen reducir la precisión de modelos CNN preentrenados en datos sin perturbaciones. Este conflicto entre precisión estándar y robustez adversarial se conoce como trade-off y refleja la necesidad de encontrar un equilibrio que mantenga un buen desempeño en datos limpios al mismo tiempo que se incrementa la robustez adversarial frente a ataques adversariales.

3.4.3. TRADES

El método TRadeoff-inspired Adversarial DEfense via Surrogate-loss mini-mization (TRADES), originalmente propuesto por (H. Zhang et al., 2019), es una estrategia basada en el preentrenamiento y derivada del adversarial training (AT). Su objetivo es mejorar la robustez adversarial de un modelo sin sacrificar excesivamente la precisión en datos limpios. Para ello, utiliza una función de pérdida compuesta por dos términos: la pérdida de clasificación estándar que busca mantener un buen rendimiento en muestras limpias y el término de robustez adversarial basado en la divergencia de Kullback-Leibler que penaliza cambios significativos entre las predicciones limpias y las adversariales; el hiperparámetro λ controla el compromiso entre ambos términos. La formulación de la pérdida es:

$$\mathcal{L}_{TRADES}(f, x, y) = \underbrace{\mathcal{L}_{CE}(f(x), y)}_{\text{Precición normal}} + \lambda \underbrace{KL(f(x)||f(x+\delta))}_{\text{Penalización de la robustez}} \tag{3.2}$$

Tal y como se muestra en la Figura 3.3, el funcionamiento de los métodos de defensa AT y TRADES junto a la función de pérdida controlada por el hiperparámetro.

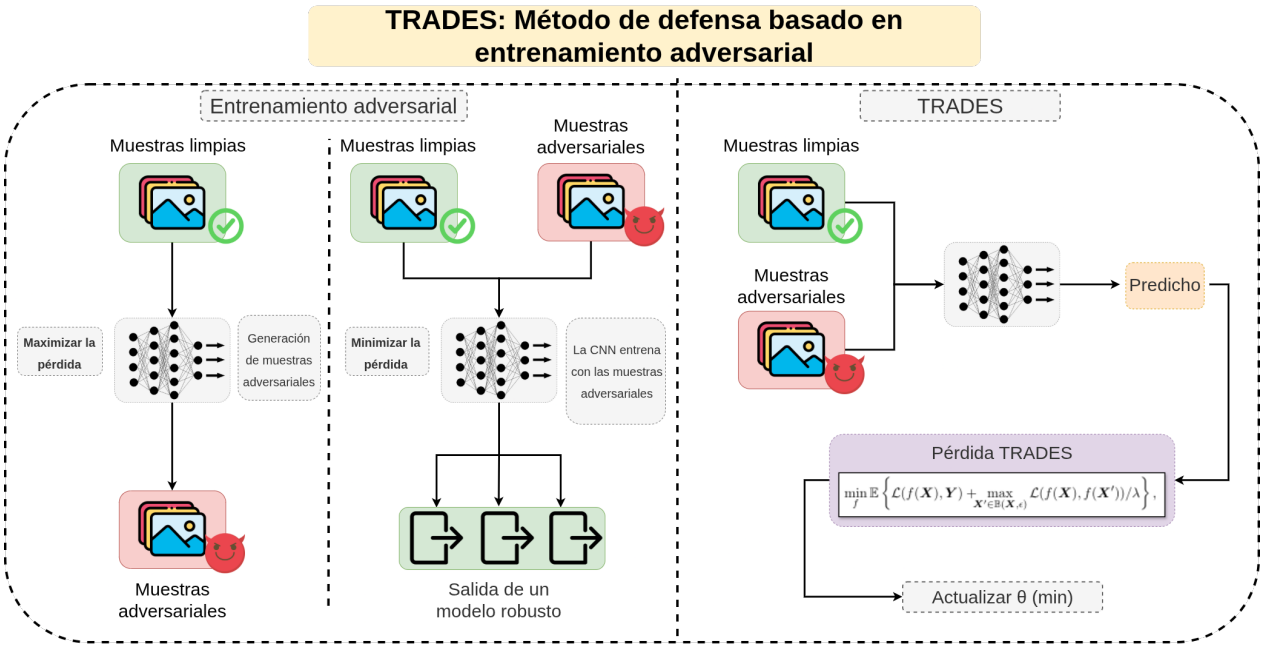

Figura 3.3: Método TRADES basado en Adversarial Training para el manejo del problema de trade-off. En el primer apartado se muestra el funcionamiento del Adversarial Training aplicando el problema de min-max para la creación de ejemplos adversariales y posterior entrenamiento. Adicionalmente, se observa el funcionamiento del método TRADES con el manejo de la predicción según el input aplicado a su función de pérdida y, finalmente, la actualización del mismo.

3.5. Dataset

Para los fines de esta investigación se utilizarán datasets compuestos por conjuntos de imágenes, los cuales serán procesados por modelos CNN preentrenados. La selección de estos datasets se basará en aquellos más utilizados en investigaciones recientes. Siendo los siguientes:

- **CIFAR-10**: Conjunto compuesto por 60.000 imágenes a color de tamaño 32×32 píxeles, distribuidas en 10 clases con 6.000 imágenes por clase. Se divide convencionalmente en 50.000 imágenes para entrenamiento y 10.000 para pruebas.
- **MNIST**: Conjunto formado por 70.000 imágenes en escala de grises de tamaño 28×28 píxeles, que representan dígitos manuscritos del 0 al 9. Se compone de 60.000 imágenes para la fase de entrenamiento y 10.000 para pruebas.

Esto tiene como objetivo facilitar la comparación de resultados con investigaciones relacionadas con los métodos de defensa y asegurar la coherencia y estandarización de los datos utilizados en el entrenamiento.

4. ESTADO DEL ARTE

Uno de los problemas más relevantes en el área del aprendizaje profundo son los ataques adversariales (Madry et al., 2018). Este fenómeno afecta a múltiples áreas de aplicación del aprendizaje profundo y compromete la robustez adversarial de diversas arquitecturas, como CNNs, modelos de lenguaje de gran escala (LLM) y otras variantes modernas (Dhamija and Bansal, 2024). Su impacto es especialmente crítico en sistemas de alta sensibilidad, tales como la conducción autónoma, el ámbito de la salud o los vehículos aéreos no tripulados.

En respuesta a estas amenazas, se han propuesto técnicas de defensa que abarcan distintas etapas del ciclo de vida del aprendizaje automático (Wu et al., 2023). El objetivo de estas defensas es dotar a los modelos de mayor robustez adversarial frente a una amplia variedad de tipos de ataques (Dhamija and Bansal, 2024). Las cuales, además de diferenciarse por tipos, se clasifican considerando el nivel de conocimiento que el atacante posee sobre el modelo.

En este contexto, en los últimos años se ha profundizado el desarrollo en el ámbito del aprendizaje profundo, particularmente en las CNN. Las cuales, como se mencionó anteriormente, cuentan con diversos enfoques y una amplia variedad de técnicas de defensa asociadas; sin embargo, todas comparten una limitación en común. Al aplicar métodos de defensa en modelos CNN estándar, se logra incrementar la robustez adversarial, pero a costa de una disminución inevitable en la precisión estándar bajo condiciones normales. Este fenómeno, conocido como problema de trade-off entre robustez adversarial y precisión estándar (Tsipras et al., 2019), afecta directamente a los modelos de visión por computadora, produciendo sistemas desbalanceados e ineficaces para su aplicación en escenarios reales. Así, se configura uno de los principales desafíos en la implementación práctica de defensas.

Inicialmente, las apariciones del fenómeno de trade-off datan del año 2014, cuando (Szegedy et al., 2014) evidenciaron que, si bien la generación de estos ejemplos podía mejorar la capacidad de los modelos para resistir perturbaciones pequeñas y, en consecuencia, incrementar su robustez adversarial, aún no se comprendían con claridad sus efectos sobre el rendimiento general ni sobre la precisión estándar en muestras limpias. Posteriormente, diversos trabajos introdujeron técnicas de defensa como el Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) y el AT (Madry et al., 2018). Estos evidenciaron los primeros indicios empíricos del denominado problema de trade-off, manifestado en una reducción significativa de la precisión estándar al priorizar la robustez adversarial en el proceso de entrenamiento.

A partir de estos hallazgos surgieron las primeras hipótesis y esfuerzos de formalización teórica del trade-off, entre ellos (Tsipras et al., 2019), donde se planteó que la robustez adversarial y la precisión estándar constituían atributos mutuamente excluyentes. En este contexto, se buscaron soluciones que minimizaran el efecto de este problema, emergiendo métodos de defensa orientados específicamente a equilibrar la tensión entre precisión estándar y robustez adversarial, destacando enfoques como TRADES (H. Zhang et al., 2019), Randomized Smoothing (Cohen et al., 2019) o Robust Self-Training (RST) (Carmon et al., 2019).

Más recientemente, las investigaciones han profundizado en este concepto, cuestionando si el trade-off es realmente una limitación intrínseca de los modelos o si, por el contrario, se encuentra condicionado por las metodologías de entrenamiento adversarial actualmente utilizadas (J. Zhang et al., 2020). Esta última perspectiva ha abierto nuevas líneas de investigación centradas en el diseño de estrategias en busca del mayor equilibrio entre robustez adversarial y precisión estándar. Desarrollando métodos como el propuesto por Kamath et al., 2021, llamado Curriculum-based Spatial-Adversarial Robustness training for Pareto-Optimality (CuSP), una técnica aplicada en el entrenamiento basada en el curriculum learning. Esta estrategia consiste en entrenar progresivamente con transformaciones espaciales y perturbaciones adversarias cada vez más complejas, de modo que la red aprenda primero tareas más simples e incremente gradualmente la dificultad, alcanzando un mejor equilibrio entre ambas métricas. En la misma línea, J. Zhang et al., 2020 propone el método Friendly Adversarial Training (FAT), que, a diferencia del entrenamiento adversario clásico, FAT no utiliza ejemplos máximamente adversarios que maximizan la pérdida, sino ejemplos más "amigables", que generan errores confiados en el modelo sin llegar al límite más agresivo del ataque. Esto se logra aplicando el ataque Projected Gradient Descent (PGD) con detención temprana, evitando alcanzar perturbaciones demasiado fuertes.

Por otro lado, Anderson and Sojoudi, 2022 propone la técnica de Locally Biased Randomized Smoothing (LBRS) para el problema de trade-off aplicada en el post-entrenamiento que, en lugar de usar distribuciones uniformes, esta técnica aprende a ajustar la distribución de ruido en función de los datos y la geometría de la frontera de decisión, eliminando la dependencia de la suposición de que más suavidad siempre significa más robustez adversarial. Un enfoque similar es el planteado por Bai et al., 2024, quien propone una técnica de defensa basada en la fase de inferencia, en la cual, por medio de combinar características de técnicas de suavizado clásico, como la regularización local de Lipschitz, con un enfoque adaptativo que evalúa la sensibilidad de la entrada del modelo para determinar cuánto suavizado aplicar en cada caso. Esto realizado sin la necesidad de reentrenamiento del modelo, simplemente modificando e interpolando entre funciones de clasificación.

A partir de estas observaciones, esta investigación propone una estrategia alternativa basada en TRADES, en la cual el parámetro interno λ se reemplaza por una función dinámica dependiente de cada muestra. El propósito de esta modificación es explorar si un ajuste adaptativo puede influir en el equilibrio entre precisión estándar y robustez adversarial, evitando así la dependencia de un hiperparámetro fijo reportada en estudios recientes. La propuesta no asume de antemano una mejora respecto a las defensas existentes; más bien, busca evaluar empíricamente si la adaptación de $\lambda(x)$ ofrece ventajas o revela nuevas limitaciones en escenarios adversariales complejos, en particular considerando las inestabilidades observadas en configuraciones tradicionales de TRADES.

En consecuencia, este trabajo explora una variante de TRADES en la cual el parámetro λ se redefine como una función dinámica dependiente de la información de cada muestra. El objetivo de esta propuesta es analizar si un ajuste adaptativo puede influir en el comportamiento del trade-off o en la estabilidad del entrenamiento frente a perturbaciones adversariales. Más que asumir una mejora respecto de los métodos consolidados, el enfoque busca evaluar empíricamente si la incorporación de $\lambda(x)$ aporta beneficios, limitaciones o comportamientos distintos a los observados en TRADES tradicional.

5. PROPUESTA DE SOLUCIÓN

En este trabajo se propone redefinir el parámetro λ de TRADES como una función dinámica dependiente de cada muestra, denominada $\lambda(x)$. Esta formulación incorpora información sobre la incertidumbre y la sensibilidad local del modelo, permitiendo ajustar la ponderación del término adversarial en función de las características específicas de cada entrada. La entropía de la predicción se utiliza para capturar la incertidumbre del modelo, mientras que la sensibilidad adversarial se aproxima mediante el gradiente de la divergencia KL entre las predicciones limpias y perturbadas. En conjunto, estos elementos permiten explorar cómo un ajuste adaptativo puede modificar el comportamiento del método original sin depender de un hiperparámetro fijo. La función de λ dinámico propuesta se define como:

$$\lambda(x) = \alpha \cdot \underbrace{\mathcal{H}(f(x))}_{\text{Entropía de predicción}} + \beta \cdot \underbrace{\|\nabla_{x+\delta}\, KL(f(x) \,\|\, f(x+\delta))\|_2}_{\text{Sensibilidad adversarial}} \tag{5.1}$$

Donde α y β son hiperparámetros fijos que controlan la contribución relativa de la incertidumbre y de la defensa; por un lado, α controla cuánto influye la incertidumbre de predicción; si este valor es alto, se prioriza la robustez; en caso contrario, se comporta más como TRADES estándar. En el caso de β, controla cuánto influye la sensibilidad adversarial local; si este valor es alto, prioriza una defensa fuerte local; en caso contrario, la defensa depende más de la entropía que de la sensibilidad, mientras que $\mathcal{H}(f(x))$ representa la incertidumbre del vector probabilidad sobre la muestra x:

$$\mathcal{H}(f(x)) = -\sum_{c=1}^{C} f(x)_c \log f(x)_c \tag{5.2}$$

Donde $f(x)_c$ es la probabilidad asignada a la clase c, C es el número total de clases y $\log f(x)_c$ corresponde al logaritmo natural de dicha probabilidad; el signo negativo garantiza que la entropía sea positiva, dado que $\log f(x)_c \leq 0 \; \forall \; f(x)_c \in \;]0, 1]$. Este término mide la incertidumbre epistemológica, o sea, cuando el modelo no está seguro de la clase, las probabilidades de las clases se distribuyen de forma más uniforme, aumentando la entropía. Por lo tanto, una entropía alta aumenta $\lambda(x)$, reforzando la regularización adversarial en muestras adversariales. Por otro lado, el término

$$\|\nabla_{x+\delta}\, KL(f(x) \,\|\, f(x+\delta))\|_2 \tag{5.3}$$

Mide la sensibilidad adversarial local, es decir, qué tan rápido varía la divergencia entre las distribuciones de salida ante pequeñas perturbaciones en la muestra. Los términos son x es la muestra natural, $x + \delta$ es la muestra adversarial, δ es la perturbación limitada por una condición $||\delta||_\infty \leq \epsilon$, $f(x)$ es el vector de salida del modelo para una muestra natural, $f(x+\delta)$ es el vector de salida del modelo para una muestra adversarial, $KL(f(x) \,\|\, f(x+\delta))$ es la divergencia de Kullback-Leibler entre las dos distribuciones, esta mide cuanto cambia la distribución de predicciones del modelo cuando se altera la muestra, luego esta $\nabla_{x+\delta}$ que es la gradiente del KL respecto a la muestra adversarial; mide que tan sensible es el KL respecto a cambios en $x + \delta$, finalmente, $|| \cdot ||_2$

que es la norma L_2 de la gradiente, donde se mide la magnitud total de esa sensibilidad, si esta norma es grande significa que pequeños cambios en la muestra producen grandes variaciones en la divergencia, lo que indica vulnerabilidad adversarial local.

La formulación de $\lambda(x)$ se inspira en dos líneas de trabajo previas que abordan el equilibrio entre precisión estándar y robustez adversarial desde perspectivas complementarias. Primero, Stutz et al., 2020 introducen el entrenamiento adversarial calibrado por confianza (CCAT), donde se reduce la confianza del modelo en muestras adversariales para lograr una regularización adaptativa. Este principio motiva el uso de la entropía de predicción $\mathcal{H}(f(x))$ como medida de incertidumbre, de modo que las muestras con alta entropía reciban mayor ponderación adversarial. Por otro lado, Ross and Doshi-Velez, 2018 proponen la regularización del gradiente de la muestra, demostrando que limitar la magnitud $|\nabla_x \mathcal{L}|$ aumenta la robustez y la interpretabilidad del modelo. Este hallazgo inspira el componente de sensibilidad adversarial en $\lambda(x)$, que amplifica la penalización en regiones donde pequeñas perturbaciones provocan grandes variaciones en la salida del modelo. En conjunto, ambos enfoques justifican una $\lambda(x)$ que se adapta localmente a la incertidumbre y la vulnerabilidad de cada muestra. Una vez definida $\lambda(x)$, esta se incorpora directamente en la función de pérdida de TRADES, dando lugar a la siguiente formulación:

$$\mathcal{L}_{D\text{-}TRADES}(f, x, y) = \underbrace{\mathcal{L}_{CE}(f(x), y)}_{\text{Precición normal de TRADES}} + \underbrace{\lambda(x) \cdot KL(f(x) \,\|\, f(x+\delta))}_{\text{Penalización de la robustez dinámico}} \tag{5.4}$$

De esta forma, D-TRADES conserva la estructura teórica de TRADES, pero introduce un ajuste adaptativo y local del término de regularización, eliminando la necesidad de definir un λ global y busca mitigar la inestabilidad reportada por Li et al., 2024. En la figura 5.1 se muestra el diagrama general del método propuesto, donde el modelo recibe muestras limpias y adversariales, calcula las salidas $f(x)$ y $f(x+\delta)$, y ajusta dinámicamente el parámetro de regularización $\lambda(x)$ en función de la entropía de predicción y la sensibilidad adversarial. Este $\lambda(x)$ se incorpora en la pérdida $\mathcal{L}_{D\text{-}TRADES}$ para equilibrar la precisión estándar y la robustez durante la retropropagación, preservando la estructura teórica del TRADES original. La figura 5.1 muestra un esquema general de la propuesta.

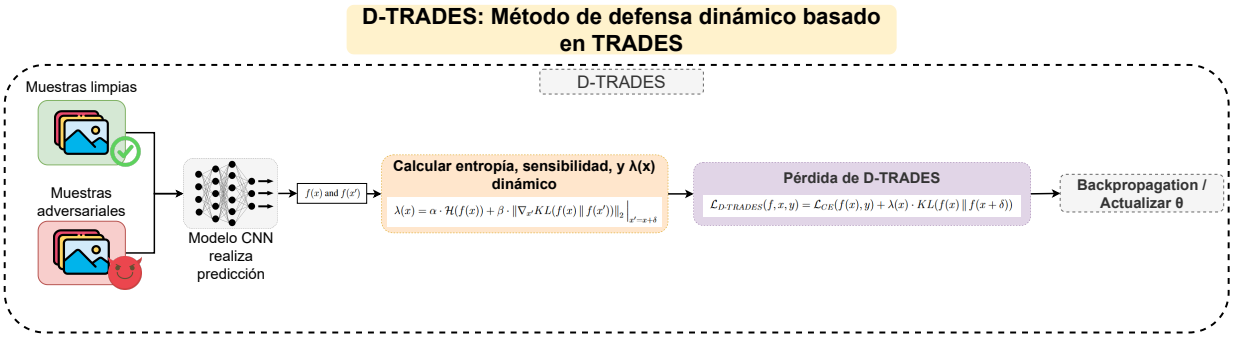

Figura 5.1: Diagrama general del método D-TRADES

Finalmente, el Algoritmo 5.1 resume la implementación práctica del método propuesto. Este mantiene la estructura general del entrenamiento adversarial utilizado en TRADES, pero incorpora el cálculo muestra a muestra de $\lambda(x)$. Durante cada iteración se generan ejemplos adversariales mediante PGD, se calculan las predicciones limpias y perturbadas, se obtienen las pérdidas naturales y robustas, y posteriormente se evalúan la entropía y sensibilidad para cada elemento del batch. Tras ello, se normalizan ambos términos y se combinan linealmente mediante los coeficientes α y β, obteniendo un $\lambda(x)$ adaptativo que pondera de manera diferenciada el término robusto de la pérdida.

Algoritmo 5.1 Función de pérdida D-TRADES

Require: Modelo f_θ, muestras limpias x, etiquetas y, tamaño de pasos η, perturbación ϵ, número de pasos del ataque T, hiperparámetros α, β

Ensure: Pérdida total $\mathcal{L}_{D\text{-}TRADES}$

Inicializar $x' \leftarrow x + 0.001 \cdot \mathcal{N}(0, 1)$

for $t = 1$ **to** T **do**

Actualizar x' mediante PGD sobre la divergencia $KL(f_\theta(x'), f_\theta(x))$

Proyectar x' dentro de la bola ϵ bajo la norma seleccionada (L_∞ o L_2)

end for

Calcular predicciones para muestra limpia: $p \leftarrow f_\theta(x)$

Calcular predicciones para muestra adversarial: $q \leftarrow f_\theta(x')$

Calcular pérdidas limpia: $\mathcal{L}_{nat} \leftarrow \text{CE}(p, y)$

Calcular Kullback-Leibler de las muestras: $KL_i \leftarrow KL(p_i \,\|\, q_i)$

Calcular entropía de predicción: $\mathcal{H}_i \leftarrow -\sum_{c=1}^{C} p_{i,c} \log(p_{i,c} + \texttt{EPS})$

Calcular sensibilidad adversarial local: $s_i \leftarrow \|\nabla_{x'_i} KL(p_i \,\|\, q_i)\|_2$

if normalizar términos **then**

Normalizar $\mathcal{H}_i$ y s_i al rango $[0, 1]$

end if

Construir parámetro dinámico: $\lambda_i \leftarrow \alpha \cdot \mathcal{H}_i + \beta \cdot s_i$

Calcular pérdida robusta ponderada: $\mathcal{L}_{rob} \leftarrow \frac{1}{B} \sum_{i=1}^{B} \lambda_i \cdot KL_i$

Calcular pérdida total: $\mathcal{L}_{D\text{-}TRADES} \leftarrow \mathcal{L}_{nat} + \mathcal{L}_{rob}$

Return $\mathcal{L}_{D\text{-}TRADES}$

6. EXPERIMENTOS

Los experimentos se diseñaron para evaluar la efectividad del método propuesto D-TRADES en comparación con defensas adversariales ampliamente estudiadas en la literatura, como TRADES, MART y FAAL. El objetivo principal es analizar el impacto de la función dinámica $\lambda(x)$ tanto en la precisión estándar como en la robustez adversarial, además de caracterizar el trade-off generado por cada defensa. Para ello, se siguió un protocolo experimental controlado, manteniendo parámetros consistentes entre las distintas defensas y arquitecturas empleadas. Las evaluaciones se realizaron utilizando los datasets MNIST y CIFAR-10, junto con las arquitecturas ResNet-18 y VGG-16, seleccionadas debido a su uso extendido y su relevancia en estudios de robustez adversarial.

6.1. Ambiente de implementación

En esta sección se describe el entorno utilizado para desarrollar y evaluar la propuesta D-TRADES, incluyendo las herramientas de software empleadas para la implementación y el hardware utilizado para el entrenamiento de los modelos. Este ambiente garantiza condiciones experimentales consistentes y reproducibles, permitiendo comparar de manera justa las defensas adversariales consideradas en el estudio.

6.1.1. Software

La implementación de la propuesta D-TRADES, junto con la comparación frente a las técnicas de defensa mencionadas en la línea base, se realizó utilizando las siguientes herramientas:

- **PyTorch**: Biblioteca de aprendizaje automático que proporciona las funcionalidades necesarias para definir y entrenar las arquitecturas ResNet y VGG, además de ofrecer herramientas para el preprocesamiento de datos, la construcción de los ciclos de entrenamiento y la evaluación de los modelos.
- **Google Colab**: Entorno basado en Jupyter Notebook que permitió realizar las evaluaciones iniciales, programar la propuesta y comparar los resultados obtenidos, aprovechando la GPU Nvidia Tesla T4 disponible en la plataforma.
- **Jupyter Notebook**: Entorno interactivo alojado en la máquina utilizada, donde se llevaron a cabo los entrenamientos de los modelos CNN, aplicando los métodos de defensa, los ataques adversariales y los datasets considerados en el estudio.
- **GitHub**: Plataforma en la nube utilizada para la gestión de versiones, el trabajo colaborativo y el almacenamiento del código desarrollado durante el proyecto.

6.1.2. Hardware

El entrenamiento de los modelos ResNet-18 y VGG-16, junto con los métodos de defensa TRADES, MART, FAAL y la propuesta D-TRADES, aplicados sobre los datasets MNIST y CIFAR-10, se llevó a cabo utilizando el siguiente entorno de hardware:

- **CPU**: AMD Ryzen 5 7600X.
- **GPU**: NVIDIA RTX 4060 con 8 GB de memoria GDDR6.
- **RAM**: 16 GB DDR5 a 5600 MHz.
- **Almacenamiento**: SSD NVMe M.2 de 1 TB.

6.2. Métricas de Evaluación

Para evaluar y comparar las defensas, se emplearon cuatro métricas fundamentales, utilizadas en la literatura para los métodos de defensa adversarial:

- **Robust Accuracy (RA)**: Mide la proporción de muestras correctamente clasificadas después de aplicar un ataque adversarial. Un valor alto indica mayor robustez adversarial. En la siguiente ecuación se muestra la formula de RA (Sun et al., 2023):

$$RA = \frac{1}{N} \cdot \sum_{i=1}^{N} 1[f(x_i + \delta_i) = y_i] \tag{6.1}$$

- **Natural Accuracy (NA)**: Corresponde a la precisión estándar del modelo sobre las muestras limpias. Permite evaluar el impacto del entrenamiento adversarial sobre el rendimiento estándar. En la siguiente ecuación se muestra la fórmula de NA (Sun et al., 2023):

$$NA = \frac{1}{N} \cdot \sum_{i=1}^{N} 1[f(x_i) = y_i] \tag{6.2}$$

- **Robustness Drop (RD)**: Cuantifica la pérdida de desempeño al pasar de datos limpios a adversariales. Un RD bajo indica un trade-off más favorable. La siguiente ecuación muestra su definición (Sun et al., 2023):

$$RD = NA - RA \tag{6.3}$$

- **Attack Success Rate (ASR)**: Es la proporción de muestras para las cuales el ataque logra cambiar la predicción correcta. En la siguiente ecuación se muestra la fórmula de ASR (Sun et al., 2023):

$$ASR = \frac{1}{N} \cdot \sum_{i=1}^{N} 1[f(x_i + \delta_i) \neq y_i] = 1 - RA \tag{6.4}$$

Estas métricas permiten evaluar tanto el rendimiento sobre las muestras limpias como la estabilidad del modelo bajo muestras adversariales, proporcionando una visión completa del trade-off de cada defensa.

6.3. Métodos a comparar

Para establecer un punto de comparación sólido, se consideran como líneas base cuatro métodos de entrenamiento adversarial ampliamente reconocidos en la literatura: (1) Entrenamiento estándar, (2) TRADES (H. Zhang et al., 2019), (3) MART (Wang et al., 2020), y (4) FAAL (Y. Zhang

et al., 2024). Según la literatura reciente, estos métodos representan los enfoques de defensa más efectivos para abordar el trade-off entre robustez y precisión, equilibrando la capacidad de generalización en muestras limpias con la resistencia ante perturbaciones adversariales.

6.4. Obtención de α y β para D-TRADES

Para la elección de los valores de α y β, se procesó una selección de estos hiperparámetros utilizados en la función dinámica $\lambda(x)$ de D-TRADES. El objetivo es identificar una configuración que ofrezca un equilibrio sólido entre precisión estándar y robustez adversarial. Para ello se evaluaron dos combinaciones de hiperparámetros consideradas razonables desde el punto de vista teórico:

- **Optimizador**: SGD con momentum 0.9
- **Batch size**: 128
- **MNIST**: 50 épocas, learning rate 0.01, $\epsilon = 0.3$, PGD^{20} y step size 0.01.
- **CIFAR-10**: 100 épocas, learning rate 0.1 con reducción en las épocas 90 y 100, weight decay 2×10^{-4}, $\epsilon = 0.031$, PGD^{10} y step size 0.007.
- **Parámetros específicos**:
  - **D-TRADES**:
    - $\alpha = 0.5,\ \beta = 0.5$
    - $\alpha = 1.0,\ \beta = 1.0$

Para ambas combinaciones de hiperparámetros se generaron ejemplos adversariales mediante PGD con el fin de monitorear adecuadamente el proceso de entrenamiento, permitiendo evaluar métricas como la robustez adversarial, precisión estándar, robustness drop o attack success rate.

Los resultados mostrados en las Tablas 6.1 y 6.2 manifiestan que el par de hiperparámetros $\alpha = 1.0$ y $\beta = 1.0$ proporciona un equilibrio más estable en comparación con la contraparte con valores de 0.5, en especial con el dataset CIFAR-10, donde se obtiene en la precisión estándar de 86.63% en ResNet-18 y 85.70% en VGG-16, y obtiene sistemáticamente mejores resultados frente a FSGM, PDG^{20} y AutoAttack$_\infty$, superando de manera amplia la contraparte de $\alpha = 0.5$ y $\beta = 0.5$. La diferencia es especialmente notable bajo ataques PGD^{20} y AutoAttack$_\infty$, donde $\alpha = 1.0$ y $\beta = 1.0$ obtienen desde un 20% a 25% más de robustez adversarial en ResNet-18 y un 23% a 28% más en VGG-16 para el dataset de CIFAR-10.

Tabla 6.1: Resultados de RA y NA del método D-TRADES evaluado con dos configuraciones distintas de hiperparámetros en los datasets CIFAR-10 y MNIST, utilizando la arquitectura ResNet-18.

	MNIST				CIFAR-10			
Defense	Natural	FGSM	PGD^{20}	AutoAttack$_\infty$	Natural	FGSM	PGD^{20}	AutoAttack$_\infty$
D-TRADES $\alpha = 0.5$ $\beta = 0.5$	99.35	99.12	**97.27**	98.90	85.28	40.60	26.90	14.20
D-TRADES $\alpha = 1.0$ $\beta = 1.0$	**99.45**	**99.30**	81.56	**98.90**	**86.63**	**52.90**	**45.39**	**40.00**

Tabla 6.2: Resultados de RA y NA del método D-TRADES evaluado con dos configuraciones de hiperparámetros en los datasets CIFAR-10 y MNIST, utilizando la arquitectura VGG-16.

	MNIST				CIFAR-10			
Defense	Natural	FGSM	PGD^{20}	$AutoAttack_\infty$	Natural	FGSM	PGD^{20}	$AutoAttack_\infty$
D-TRADES $\alpha = 0.5$ $\beta = 0.5$	**99.54**	**99.27**	**96.95**	**99.10**	84.75	33.59	16.47	3.6
D-TRADES $\alpha = 1.0$ $\beta = 1.0$	99.48	99.26	92.81	98.70	**85.70**	**47.56**	**40.15**	**32.50**

En resumen, Los resultados mostraron que la combinación $\alpha = 1.0$ y $\beta = 1.0$ presentó un comportamiento más estable y competitivo en comparación con $\alpha = 0.5$ y $\beta = 0.5$, particularmente en CIFAR-10. Esta configuración obtuvo mayor precisión natural y niveles de robustez relativamente superiores frente a los ataques evaluados, por lo que fue seleccionada para las siguientes etapas de experimentación.

6.5. Comparación con otros métodos de defensa adversarial

Una vez definida la configuración óptima de D-TRADES, se procedió a su comparación con los métodos de defensa de la literatura. Todos los métodos se entrenaron bajo un protocolo uniforme para asegurar una evaluación equitativa:

- **Optimizador**: SGD con momentum 0.9
- **Batch size**: 128
- **MNIST**: 50 épocas, learning rate 0.01, $\epsilon = 0.3$, PGD^{20} y step size 0.01.
- **CIFAR-10**: 100 épocas, learning rate 0.1 con reducción en las épocas 90 y 100, weight decay 2×10^{-4}, $\epsilon = 0.031$, PGD^{10} y step size 0.007
- **Parámetros específicos**:
  - **TRADES**: $\beta = 4$
  - **MART**: $\beta = 5$
  - **FAAL**: $r = 0.1$
  - **D-TRADES**: $\alpha = 1.0$, $\beta = 1.0$

Para todas las defensas se generaron ejemplos adversariales mediante PGD con el fin de monitorear adecuadamente el proceso de entrenamiento. A continuación se presentan los resultados para cada métrica evaluada.

6.5.1. Métrica precisión estándar y robustez adversarial

Los resultados de las Tablas 6.3 y 6.4 muestran diferencias entre los distintos métodos de defensa frente a ataques adversariales. En términos de precisión estándar, D-TRADES obtiene consistentemente los mejores resultados entre las defensas adversariales evaluadas. En CIFAR-10, D-TRADES alcanza 86.63% con ResNet-18, lo que representa una diferencia de $+2.65$ puntos por-

centuales respecto al segundo mejor método, que es TRADES con 83.98%. Para VGG-16 ocurre algo similar, donde D-TRADES logra tener 85.70%, superando al segundo mejor, siendo de nuevo TRADES con 81.62%, teniendo una diferencia de $+4.08$ puntos porcentuales. En MNIST, todas las defensas mantienen rendimientos estables y dentro del rango superior, siendo mayores al 99.00% para todos los métodos de defensa.

En contraste, su robustez adversarial es la más baja entre las defensas a comparar, especialmente en CIFAR-10, donde la tendencia es más marcada. Con Resnet-18, D-TRADES obtiene 52.90% en FSGM, quedando -6.51 puntos porcentuales por debajo del mejor valor, que es MART con 59.41%; en PGD^{20} alcanza 45.39%, es decir, -8.28 puntos menos que MART, que obtuvo 53.67%; y en AutoAttack$_\infty$ obtiene 40.00%, quedando -6.80 puntos bajo TRADES 46.80%. Estos patrones se repiten en VGG-16, donde D-TRADES también registra los valores más bajos en las tres métricas adversariales. En MNIST, aunque D-TRADES presenta resultados consistentes en FGSM y PGD^{20}, también se ubica por debajo de FAAL y MART. En conjunto, los resultados indican que, pese a su mayor precisión estándar, D-TRADES muestra la menor robustez adversarial frente a ataques adversariales entre los métodos comparados.

Tabla 6.3: Tabla de resultados de la RA y NA de los métodos de defensa con datasets de CIFAR-10 y MNIST usando la arquitectura de ResNet-18.

	MNIST				CIFAR-10			
Defense	Natural	FGSM	PGD^{20}	AutoAttack$_\infty$	Natural	FGSM	PGD^{20}	AutoAttack$_\infty$
Standard	**99.58**	98.23	9.72	95.90	**92.53**	16.85	0	0
TRADES	99.43	99.24	66.43	98.70	83.98	58.32	52.43	**46.80**
MART	99.56	**99.46**	79.91	99.20	82.53	**59.41**	**53.67**	46.10
FAAL	99.31	99.08	**98.41**	**99.30**	81.03	55.99	51.35	44.60
D-TRADES	99.45	99.30	81.56	98.90	86.63	52.90	45.39	40.00

Tabla 6.4: Tabla de resultados de la NA y RA de los métodos de defensa con datasets de CIFAR-10 y MNIST usando la arquitectura de VGG-16

	MNIST				CIFAR-10			
Defense	Natural	FGSM	PGD^{20}	AutoAttack$_\infty$	Natural	FGSM	PGD^{20}	AutoAttack$_\infty$
Standard	**99.61**	98.60	30.90	97.50	**92.19**	20.52	1.88	0
TRADES	99.44	99.34	78.07	**99.40**	81.62	55.15	49.44	44.30
MART	99.54	**99.40**	94.77	99.70	79,12	**56.50**	**51.97**	43.30
FAAL	99.26	98.97	**97.52**	98.70	79.01	54.68	50.17	**44.60**
D-TRADES	99.48	99.26	92.81	98.70	85.70	47.56	40.15	32.50

En conjunto, los experimentos muestran que D-TRADES ofrece la mejor precisión estándar entre los métodos de defensa evaluados, tanto en CIFAR-10 como en MNIST, superando consistentemente a TRADES, MART y FAAL. Sin embargo, esta ganancia se obtiene a costa de una menor robustez adversarial, especialmente notable en CIFAR-10, donde D-TRADES obtiene los valores

más bajos frente a FGSM, PGD20 y AutoAttack$_\infty$. Aunque en MNIST su rendimiento robusto es más estable, sigue situándose por debajo de FAAL y MART. Estos resultados confirman que D-TRADES prioriza la precisión estándar por sobre la resistencia a ataques, mostrando un trade-off menos equilibrado respecto de las demás defensas comparadas.

6.5.2. Métrica Robustness Drop

Por otro lado los resultados presentados en las Tablas 6.5 y 6.6, donde los valores obtenidos muestran que D-TRADES presenta el trade-off menos favorable entre las defensas evaluadas, tanto en ResNet-18 como en VGG-16. En CIFAR-10, D-TRADES alcanza los RD más altos en las tres métricas adversariales: 33.93% para el FSGM, 41.44% para el PGD^{20} y 40.77% para AutoAttack$_\infty$, superando ampliamente los valores de TRADES, MART, FAAL, cuyos RD se mantienen en rangos considerablemente menores. Esto implica que, aunque D-TRADES mantiene una precisión estándar superior, como lo visto en los resultados en las dos anteriores Tablas 6.3 y 6.4, su rendimiento cae drásticamente bajo ataques adversariales. En MNIST, el comportamiento es más estable, pero D-TRADES tampoco logra los RD más bajos, quedando por debajo de FAAL y MART.

Tabla 6.5: Resultados de la métrica Robustness Drop para los métodos de defensa evaluados en los datasets CIFAR-10 y MNIST, utilizando la arquitectura ResNet-18.

	MNIST			CIFAR-10		
Defense	FGSM	PGD^{20}	AutoAttack$_\infty$	FGSM	PGD^{20}	AutoAttack$_\infty$
Standard	1.35	89.86	3.5	75.68	92.53	92.80
TRADES	0.19	33.00	0.5	25.66	31.55	37.70
MART	**0.1**	19.65	0.2	**23.12**	**28.86**	**36.40**
FAAL	0.23	**0.90**	**0.1**	25.04	29.68	37.80
D-TRADES	0.15	17.89	0.5	33.93	41.44	47.77

Tabla 6.6: Resultados de la métrica Robustness Drop para los métodos de defensa evaluados en los datasets CIFAR-10 y MNIST, utilizando la arquitectura VGG-16.

	MNIST			CIFAR-10		
Defense	FGSM	PGD^{20}	AutoAttack$_\infty$	FGSM	PGD^{20}	AutoAttack$_\infty$
Standard	1.01	68.71	1.19	71.67	90.31	92.60
TRADES	**0.10**	21.37	0.2	26.47	32.18	37.30
MART	0.14	4.77	**0.1**	**22.62**	**27.15**	**35.30**
FAAL	0.29	**1.74**	0.5	24.33	29.68	37.80
D-TRADES	0.15	17.89	0.5	33.93	41.44	47.77

En definitiva, los resultados confirman que D-TRADES prioriza la precisión estándar por sobre la robustez adversarial, lo que deriva en una mayor RD y un trade-off menos equilibrado que el del resto de las defensas. Esto se refleja especialmente en CIFAR-10, donde obtiene los RD

más altos en todas las métricas, mostrando una caída de rendimiento más pronunciada bajo ataque. Incluso en MNIST, aunque más estable, D-TRADES no alcanza los RD más bajos, reafirmando su desventaja en robustez adversarial.

6.5.3. Metrica Attack Success Rate

Por ultimo cabe mencionar los resultados de las Tablas 6.7 y 6.8, estas Tablas permite verificar que los ataques realmente ejercen presión suficiente sobre los modelos para analizar su robustez adversarial. Sin embargo, algunos ataques presentan un ASR extremadamente bajos, siendo el caso del AutoAttack$_\infty$ en MNIST, cuyos valores se mantienen cercanos al cero en todas las defensas evaluadas. Dado que un ASR tan reducido implica que el ataque no logra generar perturbaciones efectivas en ese dominio, su contribución al análisis comparativo resulta limitada.

Tabla 6.7: Resultados de la métrica Attack Success Rate para los métodos de defensa evaluados en los datasets CIFAR-10 y MNIST, utilizando la arquitectura ResNet-18.

	MNIST			CIFAR-10		
Defense	FGSM	PGD^{20}	AutoAttack$_\infty$	FGSM	PGD^{20}	AutoAttack$_\infty$
Standard	1.77	90.28	3.5	83.15	100	100
TRADES	0.76	33.57	1.3	41.68	47.57	**53.20**
MART	**0.54**	20.09	0.8	**40.59**	**46.33**	53.90
FAAL	0.92	**1.59**	**0.7**	44.01	48.65	55.40
D-TRADES	0.7	18.44	1.10	47.10	54.61	60.00

Tabla 6.8: Resultados de la métrica Attack Success Rate para los métodos de defensa evaluados en los datasets CIFAR-10 y MNIST, utilizando la arquitectura VGG-16.

	MNIST			CIFAR-10		
Defense	FGSM	PGD^{20}	AutoAttack$_\infty$	FGSM	PGD^{20}	AutoAttack$_\infty$
Standard	1.40	69.10	2.5	79.48	98.12	100
TRADES	0.66	21.93	0.6	44.85	50.56	55.70
MART	**0.6**	5.23	**0.3**	**43.50**	**48.03**	56.70
FAAL	1.03	**2.48**	1.3	45.32	49.83	**55.40**
D-TRADES	0.74	7.19	0.13	52.44	59.85	67.50

En consecuencia, en las secciones anteriores no se considera la columna correspondiente a AutoAttack$_\infty$ en MNIST, ya que los valores obtenidos carecen de relevancia analítica. Como muestran las Tablas 6.7 y 6.8, su ASR se mantiene prácticamente en cero, lo que evidencia que el ataque no genera perturbaciones efectivas y, por lo tanto, aporta poco al análisis comparativo.

6.6. Discusión de Resultados

Los resultados obtenidos permiten establecer varias conclusiones sobre el comportamiento de D-TRADES y su relación con el trade-off entre precisión natural y robustez adversarial. En primer lugar, la combinación $\alpha = 1$ y $\beta = 1$ se valida como la configuración más adecuada, especialmente por su desempeño en CIFAR-10. Si bien esta configuración genera la mayor precisión natural entre todas las defensas evaluadas, también conduce a una pérdida significativa de robustez adversarial, destacándose como el método de defensa con mayores valores de RD y ASR.

Este comportamiento sugiere que la función dinámica $\lambda(x)$ tiende a reforzar la clasificación estándar más que la penalización adversarial, lo que podría estar asociado a dos factores:

1. La entropía empíricamente aumenta en muestras difíciles, induciendo un $\lambda(x)$ mayor en regiones donde el modelo es incierto, pero no necesariamente en aquellas más vulnerables a perturbaciones.
2. La sensibilidad adversarial local presenta alta varianza, lo que podría estar generando ajustes poco estables en la regularización de la pérdida robusta según la complejidad de la muestra.

En segundo lugar, se observó un comportamiento distintivo en MNIST, donde la mayoría de los métodos de defensa alcanzaron niveles de robustez adversarial que superan el 90%, algo que desaparece por completo en CIFAR-10, donde los valores son considerablemente más moderados. Esta diferencia puede asociarse a la estructura del propio dataset MNIST, cuyas clases están notablemente separadas y se representan mediante formas simples y bien definidas que requieren perturbaciones más intensas para afectar el resultado. CIFAR-10, por el contrario, incorpora variaciones de iluminación, texturas complejas, bordes difusos y ruido natural, lo que hace que el modelo sea mucho más vulnerable a pequeñas perturbaciones. Además, en MNIST es habitual que los modelos presenten sobreajuste, lo que genera fronteras de decisión amplias y difíciles de vulnerar. Este comportamiento coincide con lo reportado en trabajos previos como MART (Wang et al., 2020).

En coherencia con estos resultados, el funcionamiento irregular de AutoAttack$_\infty$ en MNIST puede relacionarse con la naturaleza del propio dataset. Al tratarse de imágenes en escala de grises con patrones simples y baja complejidad visual, existen pocas direcciones de perturbación que permitan alterar la imagen sin comprometer su semántica. Esto reduce la efectividad de algunos de los submétodos que componen AutoAttack$_\infty$, ya sea por la falta de fronteras de decisión finas, por la ausencia de información RGB o por la tendencia a converger demasiado rápido hacia soluciones que parecen óptimas. No se descarta, además, que una configuración interna poco adecuada del ataque haya contribuido al fenómeno observado.

7. CONCLUSIÓN Y TRABAJO FUTURO

En esta investigación se abordó el problema del trade-off entre precisión estándar y robustez adversarial, un desafío ampliamente reconocido en el ámbito de las defensas para modelos CNN. Las técnicas actuales, como TRADES, MART y FAAL, buscan mitigar esta tensión mediante distintos enfoques. Es sobre esta base, que se desarrolló la propuesta D-TRADES, concebida como una modificación del funcionamiento original de TRADES al reemplazar el valor fijo de λ por una formulación dependiente de los parámetros α y β, con el objetivo de regular tanto la incertidumbre de la predicción como la sensibilidad local del modelo.

A partir de este planteamiento, la fase experimental permitió comparar el desempeño de D-TRADES con los métodos existentes. Los resultados mostraron que la propuesta no supera a las defensas consolidadas en términos de robustez adversarial y presenta un comportamiento que se inclina mayormente hacia la precisión estándar, comprometiendo especialmente su rendimiento frente a ataques de mayor intensidad. Esto evidencia que el modelo aún no consigue sostener un equilibrio estable ante perturbaciones adversariales. Sin embargo, los experimentos también revelaron que, en diversos escenarios, D-TRADES logra aproximarse de manera competitiva a los métodos tradicionales, lo que indica que su estructura tiene potencial y que un ajuste más cuidadoso de los hiperparámetros podría mejorar significativamente sus resultados.

A partir de estas consideraciones, se identifican cuatro direcciones para el trabajo futuro. La primera consiste en optimizar el rendimiento del método mediante estrategias avanzadas de búsqueda de hiperparámetros, incluyendo técnicas metaheurísticas. La segunda propone explorar arquitecturas CNN diseñadas específicamente para ajustarse mejor a las dinámicas del enfoque. Asimismo, resulta relevante analizar el costo computacional asociado al entrenamiento de modelos robustos, con el fin de evaluar la viabilidad de su uso en sistemas reales, especialmente a medida que aumenta la complejidad de los modelos y defensas. Finalmente, es pertinente estudiar el comportamiento del método en escenarios prácticos, aplicándolo a imágenes industriales o médicas donde la robustez adversarial adquiere una importancia crítica.

8. PLANIFICACIÓN

En esta sección se presenta la planificación semanal para el desarrollo del seminario de título, considerando las actividades y entregas programadas a lo largo del semestre, esto mostrado en la Tabla 8.1.

Tabla 8.1: Planificación de actividades semestral

Actividad	**Meses**															
	Agosto				Septiembre				Octubre				Noviembre			
Definición e investigación de propuesta a tratar	X															
Planificación de actividades	X															
Definición de espacios de trabajo		X														
Revisión de objetivos		X														
Revisión y redacción de estado del arte e introducción		X	X													
Estructuración y redacción de informe			X													
Preparación primer avance				X												
Presentación primer avance					X											
Revisión de feedback primer avance					X	X										
Confección de propuesta						X										
Confección de marco teórico						X										
Recopilación de tecnologías a utilizar							X									
Implementación técnica de la propuesta							X	X	X	X						
Experimentación										X	X					
Evaluación de resultados												X	X	X		
Redacción de informe y presentación final														X	X	
Presentación final																X

REFERENCIAS

Anderson, B. G., & Sojoudi, S. (2022, January). Certified robustness via locally biased randomized smoothing. In R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, & M. Kochenderfer (Eds.), *Proceedings of the 4th annual learning for dynamics and control conference* (pp. 207–220, Vol. 168). PMLR. https://proceedings.mlr.press/v168/anderson22a.html

Arani, E., Sarfraz, F., & Zonooz, B. (2020). Adversarial Concurrent Training: Optimizing Robustness and Accuracy Trade-off of Deep Neural Networks. *British Machine Vision Conference*. https://www.bmvc2020-conference.com/assets/papers/0859.pdf

Bai, Y., Anderson, B. G., Kim, A., & Sojoudi, S. (2024). Improving the accuracy-robustness trade-off of classifiers via adaptive smoothing. https://arxiv.org/abs/2301.12554

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., & Liang, P. S. (2019). Unlabeled data improves adversarial robustness. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/32e0bd1497aa43e02a42f47d9d6515ad-Paper.pdf

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019, June). Certified adversarial robustness via randomized smoothing. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (pp. 1310–1320, Vol. 97). PMLR. https://proceedings.mlr.press/v97/cohen19c.html

Dhamija, L., & Bansal, U. (2024). How to defend and secure deep learning models against adversarial attacks in computer vision: A systematic review. *New Generation Computing*, *42*. https://doi.org/10.1007/s00354-024-00283-0

Gheisari, M., Ebrahimzadeh, F., Rahimi, M., Moazzamigodarzi, M., Liu, Y., Dutta Pramanik, P. K., Heravi, M. A., Mehbodniya, A., Ghaderzadeh, M., Feylizadeh, M. R., & Kosari, S. (2023). Deep learning: Applications, architectures, models, tools, and frameworks: A comprehensive survey. *CAAI Transactions on Intelligence Technology*, *8*(3), 581–606. https://doi.org/https://doi.org/10.1049/cit2.12180

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. https://arxiv.org/abs/1412.6572

Hassan, A., Hemeida, A., & Hassan, M. (2022). Image classification based deep learning: A review. *Aswan University Journal of Sciences and Technology*, *2*. https://doi.org/10.21608/aujst.2022.259887

Kamath, S., Deshpande, A., Kambhampati Venkata, S., & N Balasubramanian, V. (2021). Can we have it all? on the trade-off between spatial and adversarial robustness of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (pp. 27462–27474, Vol. 34). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/e6ff107459d435e38b54ad4c06202c33-Paper.pdf

Li, J. W., Liang, R.-W., Yeh, C.-H., Tsai, C.-C., Yu, K., Lu, C.-S., & Chen, S.-T. (2024). Adversarial robustness overestimation and instability in trades. *CoRR*, *abs/2410.07675*. https://doi.org/10.48550/arXiv.2410.07675

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv (Cornell University)*. http://arxiv.org/pdf/1706.06083.pdf

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. https://arxiv.org/abs/1511.04508

Rade, R., & Moosavi-Dezfooli, S.-M. (2021). Helper-based Adversarial Training: Reducing Excessive Margin to Achieve a Better Accuracy vs. Robustness Trade-off. *International Conference on Machine Learning*. https://openreview.net/pdf?id=BuD2LmNaU3a

Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Engineering*, *6*(3), 346–360. https://doi.org/https://doi.org/10.1016/j.eng.2019.12.012

Ross, A., & Doshi-Velez, F. (2018). Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11504

Song, J., Gao, S., Zhu, Y., & Ma, C. (2019). A survey of remote sensing image classification based on cnns. *Big Earth Data*, *3*(3), 232–254. https://doi.org/10.1080/20964471.2019.1657720

Song, Y., Kim, T., Nowozin, S., Ermon, S., & Kushman, N. (2018). Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. https://arxiv.org/abs/1710.10766

Stutz, D., Hein, M., & Schiele, B. (2020, 13–18 Jul). Confidence-calibrated adversarial training: Generalizing to unseen attacks. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (pp. 9155–9166, Vol. 119). PMLR. https://proceedings.mlr.press/v119/stutz20a.html

Sun, J., Chen, L., Xia, C., Zhang, D., Huang, R., Qiu, Z., Xiong, W., Zheng, J., & Tan, Y.-A. (2023). Canary: An adversarial robustness evaluation platform for deep learning models on image classification. *Electronics*, *12*(17). https://doi.org/10.3390/electronics12173665

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. https://arxiv.org/abs/1312.6199

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2019). Robustness may be at odds with accuracy. https://arxiv.org/abs/1805.12152

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., & Gu, Q. (2020). Improving adversarial robustness requires revisiting misclassified examples. *International Conference on Learning Representations*. https://openreview.net/forum?id=rklOg6EFwS

Wu, B., Wei, S., Zhu, M., Zheng, M., Zhu, Z., Zhang, M., Chen, H., Yuan, D., Liu, L., & Liu, Q. (2023). Defenses in adversarial machine learning: A survey. https://arxiv.org/abs/2312.08890

Xu, W., Evans, D., & Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. https://arxiv.org/abs/1704.01155

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., & Jordan, M. I. (2019, January). Theoretically principled trade-off between robustness and accuracy. https://arxiv.org/abs/1901.08573

Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., & Kankanhalli, M. (2020, July). Attacks which do not kill training make adversarial learning stronger. In H. D. III & A. Singh (Eds.),

Proceedings of the 37th international conference on machine learning (pp. 11278–11287, Vol. 119). PMLR. https://proceedings.mlr.press/v119/zhang20z.html

Zhang, Y., Zhang, T., Mu, R., Huang, X., & Ruan, W. (2024). Towards fairness-aware adversarial learning. https://arxiv.org/abs/2402.17729