ESCUELA DE INGENIERÍA INFORMÁTICA

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

D-TRADES: BALANCE ADAPTATIVO ENTRE ROBUSTEZ ADVERSARIAL Y PRECISIÓN APLICADO AL ENTRENAMIENTO EN 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

SEPTIEMBRE 2025

Dedicatoria (se incluye solo en TFT o TFG).
Una dedicatoria es una breve expresión de gratitud
o reconocimiento, generalmente personal y emotiva,
donde el autor dedica su trabajo a una o
varias personas significativas en su vida.

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 han centrado en incrementar la robustez adversarial, sin considerar adecuadamente el impacto en la precisión. En este trabajo se propone 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. Se espera que la propuesta optimice el equilibrio robustez-precisión, mantenga competitividad frente a ataques adversariales y considere métricas actuales.

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

ABSTRACT

The trade-off problem in the context of adversarial attacks arises when applying defense methods to convolutional neural networks to increase their adversarial robustness. While these methods strengthen the model's resistance, they often lead to a reduction in accuracy under standard conditions, creating an imbalance that limits their applicability in real-world scenarios. Most recent research has focused on increasing adversarial robustness without adequately considering the impact on accuracy. In this work, we propose a method aimed at optimizing the balance between standard accuracy and adversarial robustness, based on the TRADES method, in which 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 goal of achieving a more efficient balance between robustness and accuracy. A comparative study of defense methods and their associated trade-offs is conducted to analyze their effects on the overall performance of the model. This proposal is expected to optimize the robustness-accuracy balance, remain competitive against adversarial attacks, and incorporate current evaluation metrics.

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 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. En primer lugar, la precisión estándar se trata del rendimiento de un modelo con muestras naturales, es decir, entradas sin manipulaciones realizadas por terceros; en segundo lugar, 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 TRADES (TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization), es un método de entrenamiento adversarial que busca equilibrar la precisión y robustez 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, y 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 debido a gradient masking (Li et al., 2024).

En esta investigación se propone un método de defensa adversarial basado en TRADES, al que se le modifica un hiperparámetro λ fijo en la versión base por un hiperparámetro dinámico y dependiente de la entrada. Para ello, se realiza un análisis comparativo exhaustivo frente a métodos existentes, incluyendo TRADES con λ fijo, entrenamiento adversarial, MART y FAAL, utilizando las arquitecturas ResNet-18 y VGG sobre el dataset CIFAR-10. La metodología incluye tanto la sistematización del estado del arte como experimentos controlados de entrenamiento y evaluación, con métricas que permiten valorar la eficacia de la propuesta en escenarios adversariales. El objetivo es determinar si el enfoque dinámico sugerido permite alcanzar un mejor compromiso entre robustez y precisión que las estrategias actuales.

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 investi-

gación. Posteriormente, la descripción del problema profundiza en su definición, causas y posibles soluciones según la literatura. 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 sexta sección, la metodología plantea el plan de trabajo a seguir para completar los objetivos especificados. Después, 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. Finalmente, las conclusiones ofrecen una reflexión crítica sobre la propuesta, sus ventajas y limitaciones, así como posibles mejoras para futuras investigaciones.

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 y evaluar un método de trade-off basado en TRADES que utilice un parámetro de regularización dinámico dependiente de la entrada, 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 especificos

Los objetivos de este trabajo de investigación son los siguientes:

- Analizar y sistematizar el estado del arte sobre CNN 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.
- Entrenamiento de redes neuronales convolucionales (CNN) utilizando los modelos ResNet-18 y VGG, sobre los conjuntos de datos CIFAR-10 y MNIST, aplicando los siguientes métodos de defensa:
  - TRADES baseline (λ fijo)
  - D-TRADES con $\lambda(x)$ dinámico
  - Adversarial Training
  - Misclassification Aware adveRsarial Training (MART)
  - Fairness-aware adversarial learning (FAAL)
- Evaluar y comparar el rendimiento del método propuesto contra los otros modelos. 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 area de estudio como lo seria el deep learning y las redes neuronales convolucionales. Posteriormente se abarca la definicion de conceptos especificos como muestras, ataques y defensas adversariales. Finalizando breve explicacion de el metodo TRADES, trade-off y conjunto 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 ademas 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.

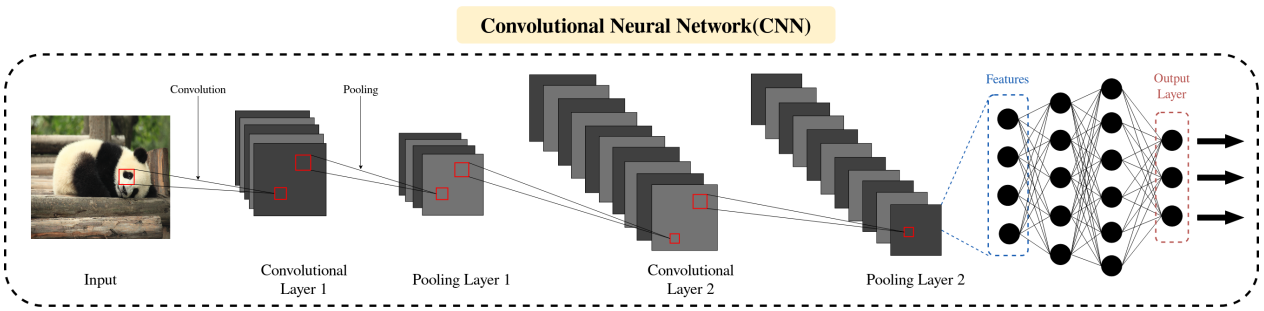

Figura 3.1: Estructura de una CNN.

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 conjuntos de datos, 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 varias 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.
- GoogLeNet: Incorpora módulos Inception que permiten incrementar profundidad y ancho de la red sin aumentar excesivamente el cómputo.
- 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. Muestras y Perturbaciones Adversariales

Las muestras adversariales se diseñan con el propósito de inducir errores en los modelos de clasificación, forzándolos a asignar etiquetas incorrectas a entradas que en realidad pertenecen a otra clase (Dhamija and Bansal, 2024). Para generarlas, se modifica una muestra original X introduciendo pequeñas perturbaciones o ruidos casi imperceptibles, obteniendo así una nueva entrada X' que, pese a ser muy similar visualmente, lleva al modelo a predecir una etiqueta diferente a la correspondiente a la muestra original.

3.4. Ataques Adversariales

Un ataque adversarial consiste en la generación de perturbaciones mínimas e imperceptibles sobre los datos de entrada, diseñadas específicamente para inducir a un modelo de aprendizaje profundo a producir predicciones incorrectas con alta confianza (Madry et al., 2018; Rade and Moosavi-Dezfooli, 2021). Como se observa en la Figura 3.2 la diferencia de clasificacion en un modelo CNN preentrenado con y sin imágenes adversariales.

3.4.1. 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 black-box, donde el atacante genera múltiple 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 mas 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:
  - Ataques dirigidos: Focalizado en inducir al modelo a clasificar una muestra adversarial dado por el atacante.
  - Ataques no dirigidos: Busca unicamente causar un error en la clasificación en el modelo.

3.4.2. 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.
  - Ensemble Surrogate
  - 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.
  - Projected Gradient Descent (PGD)
  - AutoAttack
  - FGSM
- Caja Gris: Un ataque de caja gris, el atacante tiene conocimiento unicamente de la estructura.

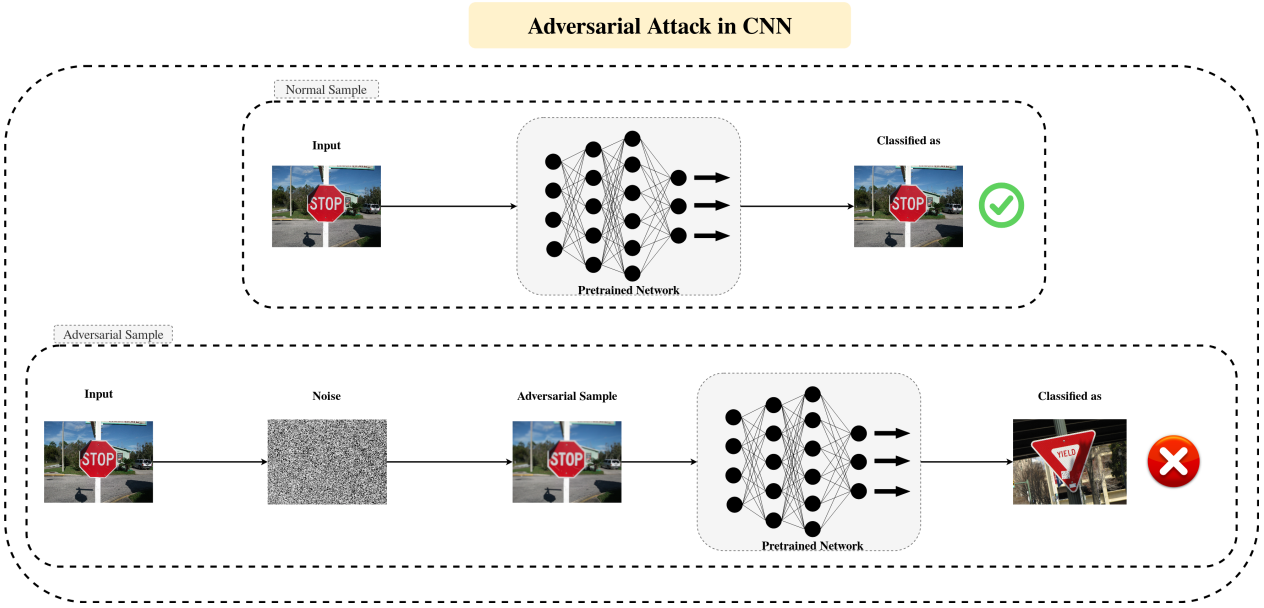

Figura 3.2: Clasificación de una CNN preentrenada frente a una entrada con y sin ruido.

3.5. 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.

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

3.5.1. Robustez Adversarial

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

3.5.2. Fenomeno "trade-off" en vision 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, suelen reducir la precisión de modelos CNN preentrenados en datos sin perturbaciones. Este conflicto entre precisión y robustez 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 frente a ataques adversariales.

3.5.3. TRADES

El método TRADES, originalmente propuesto por (H. Zhang et al., 2019), es una estrategia basada en el pre-entrenamiento 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: uno que optimiza la precisión estándar y otro que penaliza la vulnerabilidad frente a ejemplos adversariales. El equilibrio entre estos dos objetivos está controlado por el hiperparámetro λ, que permite ajustar la compensación entre precisión y robustez según las necesidades del modelo y del dominio de aplicación. 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 perdida controlada por el hiperparámetro.

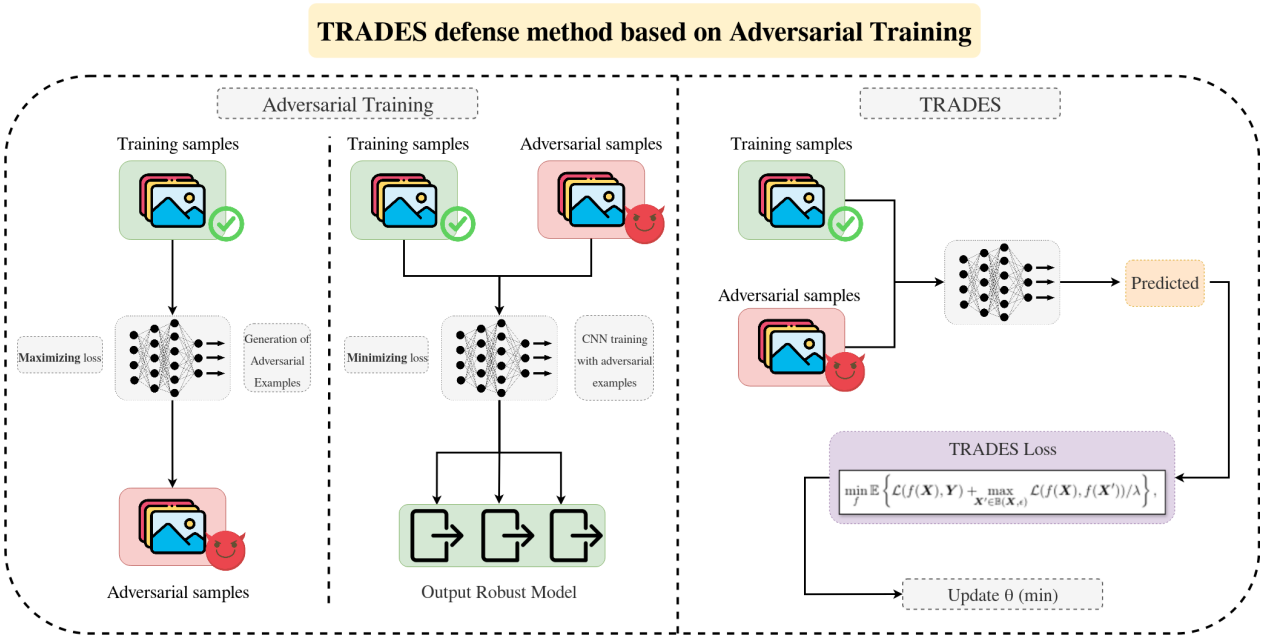

Figura 3.3: Método TRADES basado en Adversarial Training para el manejo del problema de trade-off. En el primer apartado se muestra el funcionamiento del Adversarial Training aplicando el problema de min-max para la creación de ejemplos adversariales y posterior entrenamiento. Adicionalmente se observa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

3.6. 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.

- Cifar-10
- MNIST

Esto tiene como objetivo facilitar la comparación de resultados con investigaciones relacionadas con el método TRADES 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 de diversas arquitecturas, como redes convolucionales (CNN), 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 frente a una amplia variedad de tipos de ataques (Dhamija and Bansal, 2024). Las cuales ademas de diferenciarse por tipos estas se clasifican considerando el nivel de conocimiento que el atacante posee sobre el modelo.

En este contexto que en los últimos años se ha profundizado el desarrollo en el ámbito del aprendizaje profundo, particularmente en las redes neuronales convolucionales (CNN). Las cuales como se menciono 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 bajo condiciones normales. Este fenómeno, conocido como problema de trade-off entre robustez y precisión (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, aún no se comprendían con claridad sus efectos sobre el rendimiento general ni sobre la precisión en datos naturales. Posteriormente, diversos trabajos introdujeron técnicas de defensa como el Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) y el Adversarial Training (AT) (Madry et al., 2018), los cuales 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 y la precisión 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 y robustez, destacando enfoques como TRadeoff-inspired Adversarial DEfense via Surrogate-loss mini-mization (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 y precisión. Desarrollando metodos 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 mas 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 linea 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 tecnica de LBRS (Locally Biased Randomized Smoothing) 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 que mas suavidad siempre significa mas robustez. 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 cuanto suavizado aplicar en cada caso. Esto realizado sin la necesidad de re entrenamiento del modelo, simplemente modificando e interpolando entre funciones de clasificación.

Paralelamente, se han desarrollado variantes del método TRADES que incorporan un enfoque de equidad en el entrenamiento adversarial. Entre ellas se encuentra TRades-based mIXed adversarial training (TRIX), el cual, según Medi et al., 2025, asigna dinámicamente adversarios específicos o no específicos en función de la similitud de características entre clases, mejorando la equidad en la robustez sin sacrificar la precisión estándar. Asimismo, Misclassification Aware adveRsarial Training (MART), propuesto por Wang et al., 2020, introduce una pérdida adicional que otorga mayor peso a ejemplos difíciles de clasificar, lo que permite reducir el impacto de ataques sobre las clases más vulnerables y lograr un mejor equilibrio entre robustez y precisión. Del mismo modo, Y. Zhang et al., 2024 propone Fairness-Aware Adversarial Learning (FAAL), un enfoque que prioriza la equidad por sobre la robustez tradicional, garantizando que la respuesta del modelo frente a ataques, no genere sesgos injustos entre clases o grupos, esto por medio de el uso de técnicas robustas de distribución que aseguran un rendimiento consistente incluso en el peor escenario entre categorías. Aunque los enfoques previamente descritos representan avances relevantes, esta investigación propone una nueva estrategia de trade-off que busca superar limitaciones aún presentes, cuyo aporte central esta basado en la tecnica TRADES y consiste en modificar el parámetro interno λ (lambda), transformándolo de estático a dinámico para que se adapte automáticamente según los inputs recibidos durante el entrenamiento. Esta adaptación permite asignar un

nivel de robustez proporcional a la dificultad o vulnerabilidad de cada ejemplo, mitigando sesgos internos entre clases y promoviendo equidad en la respuesta del modelo frente a ataques adversariales. Además, se espera que mejore la generalización a datos no vistos y a escenarios adversos del mundo real, contribuyendo al desarrollo de modelos de visión por computadora más confiables y equilibrados. Esto debido a investigaciones recientes de la tecnica TRADES que segun (Li et al., 2024) demuestran su alta probabilidad de inducir mascara de gradiente, lo que conduce a una percepción falsa de robustez (sobreestimación) y a una evaluación inexacta de la resistencia del modelo frente a ataques adversariales.

En síntesis, las investigaciones recientes confirman que, si bien los ataques adversariales constituyen un fenómeno relativamente frecuente en el ámbito del aprendizaje profundo, su impacto ha motivado un desarrollo sostenido de técnicas de defensa a lo largo de los últimos años. Estas defensas, aunque han permitido avances significativos, evidenciaron también un efecto colateral: el surgimiento del problema de trade-off entre robustez y precisión, el cual reveló que incrementar la resistencia frente a ataques suele implicar una disminución del rendimiento en condiciones normales. Esta tensión impulsó la aparición de métodos específicamente orientados a equilibrar ambos aspectos, con el objetivo de generar modelos más confiables y aplicables en escenarios reales. Dentro de estos enfoques, TRADES se consolidó como uno de los métodos más influyentes al sentar las bases para investigaciones posteriores; que sin embargo, estudios recientes han demostrado que no está exento de limitaciones, destacando problemas de inestabilidad en sus parámetros y fenómenos como la máscara de gradiente, comprometiendo la evaluación real de la robustez. En consecuencia, y sobre la base de estas limitaciones, este trabajo plantea una propuesta de λ (lambda) dinámico, orientada a superar dichas debilidades mediante un ajuste adaptativo que busque mejorar simultáneamente la precisión, estabilidad y robustez del modelo original, avanzando hacia arquitecturas de visión por computadora más equilibradas y confiables para su aplicación práctica.

5. METODOLOGÍA

La metodología propuesta para el desarrollo de esta investigación se estructura en torno a un enfoque experimental y comparativo, orientado a la evaluación entre defensas adversariales de la literatura. El propósito central es analizar el comportamiento de la propuesta, evaluando si es competitiva con las defensas elegidas para la comparación. El proceso metodológico se organiza en cinco etapas principales basadas en los objetivos definidos anteriormente: análisis teórico, entrenamiento e implementación de los modelos y comparación de resultados.

5.1. Análisis teórico

En este apartado se revisa una revisión sistemática del estado del arte respecto a la robustez adversarial, con énfasis en el problema de trade-off precisión–robustez y en el método TRADES. Esta revisión debe abarcar los siguientes puntos:

- Conceptos relacionados y sus definiciones
- Trade-off en las defensas adversariales
- Defensas adversariales relacionadas con TRADES

El criterio definido para elegir los papers a tener en cuenta, respecto a los conceptos, ataques y defensas adversariales, está basado en qué tanta relación tengan con nuestra propuesta y la importancia que tuvieron para la literatura y para el avance de la investigación de estos conceptos. El análisis teórico sirve de base para tener constancia del estado actual del área de investigación y tener un conocimiento más sólido del tema a tratar.

5.2. Diseño experimental

El diseño experimental considera la ejecución de tres fases:

- **Entrenamiento**: Se entrena cada modelo de red convolucional (ResNet-18 y VGG16) con la propuesta del informe y los métodos de defensa propuestos para la comparación.
- **Evaluación de los modelos**: Se someten los modelos a ataques de caja blanca (ataque a decidir) y caja negra (ataque a decidir) para evaluar su grado de robustez.
- **Comparación de resultados**: Se comparan las métricas obtenidas entre la propuesta y los métodos alternativos seleccionados.

Para la realización de las fases se utilizará el entorno Google Colab utilizando la GPU Tesla T4; como lenguaje de programación se utilizará Python 3.13 y las siguientes librerías:

- PyTorch
- NumPy

- Matplotlib

Cada fase es dependiente de la anterior, o sea, que sin realizar el entrenamiento, no se pueden evaluar los modelos y, sin esta, no se pueden comparar. Esta linealidad sirve para definir de manera clara los pasos a seguir.

5.3. Entrenamiento e implementación de los modelos

En esta etapa se implementan los modelos de defensa considerados en la investigación,

- **TRADES** (modelo base) con parámetro λ fijo.
- **D-TRADES** (propuesta), que introduce un $\lambda(x)$ dinámico dependiente de la entrada.
- **Adversarial Training**
- **MART** (Misclassification Aware Adversarial Training).
- **FAAL** (Fairness-Aware Adversarial Learning).

La implementación se basa en entrenar con arquitecturas de redes convolucionales preentrenadas (ResNet-18 y VGG16), sobre los conjuntos de datos CIFAR-10 y MNIST. El entrenamiento se lleva a cabo empleando optimización (Adam), con tasa de aprendizaje inicial (aún tengo que verlo), batch size (aún tengo que verlo) y un total de (aún tengo que verlo) épocas. Los ataques adversariales utilizados para el entrenamiento y evaluación se generan mediante (aún tengo que preguntarlo).

5.4. Comparación de resultados

Se realizará una comparación de los resultados obtenidos entre las distintas defensas. Para medir la efectividad de las defensas se emplean métricas estándar en robustez adversarial:

- **Robust Accuracy (RA)**: Mide la proporción de ejemplos correctamente clasificados incluso después de aplicar un ataque adversarial, cuanto mayor sea el valor dado, más robusta es la defensa. En la siguiente ecuación se muestra la formula de RA (Sun et al., 2023):

$$RA = \frac{1}{N} \cdot \sum_{i=1}^{N} 1[f(x_i + \delta_i) = y_i] \tag{5.1}$$

- **Natural Accuracy (NA)**: Mide la precisión del modelo sobre los datos limpios, sirve para medir el equilibrio entre precisión estándar y robustez. En la siguiente ecuación se muestra la formula de NA (Sun et al., 2023):

$$NA = \frac{1}{N} \cdot \sum_{i=1}^{N} 1[f(x_i) = y_i] \tag{5.2}$$

- **Robustness Drop (RD)**: Cuantifica cuanto cae el rendimiento del modelo al pasar de imágenes limpias a adversariales, midiendo el costo de precisión que provoca el ataque. En la siguiente ecuación se muestra la formula de RD (Sun et al., 2023):

$$RD = NA - RA \tag{5.3}$$

- **Attack Success Rate (ASR)**: Es la proporción de ejemplos en los que el ataque logra cambiar la predicción correcta, o sea, cuanto menor sea el valor, mejor defensa. En la siguiente ecuación se muestra la formula de ASR (Sun et al., 2023):

$$ASR = \frac{1}{N} \cdot \sum_{i=1}^{N} 1[f(x_i + \delta_i) \neq y_i] = 1 - RA \tag{5.4}$$

- **Average Perturbation Norm (APN)**: Evalúa el costo del ataque midiendo el tamaño promedio de las perturbaciones necesarias para engañar al modelo, o sea, cuanto mayor sea el valor, más difícil resulta atacar al modelo. En la siguiente ecuación se muestra la formula de APN (Sun et al., 2023):

$$APN = \frac{1}{N} \cdot \sum_{i=1}^{N} \min_{\delta_i} \{||\delta_i||_p f(x_i + \delta_i) \neq y_i\} \tag{5.5}$$

Cada experimento se repite tres veces con semillas aleatorias distintas para obtener promedios y desviaciones estándar confiables. Se evalúa el grado en que D-TRADES logra mejorar el equilibrio entre precisión estándar y robustez adversarial, considerando métricas enfocadas en la precisión estándar y adversarial, como también en la cantidad de veces que el modelo se equivoca clasificando.

6. PROPUESTA DE SOLUCIÓN

El método TRADES propuesto por H. Zhang et al., 2019 equilibra la precisión natural y la robustez adversarial mediante una función de pérdida compuesta por dos términos: la pérdida de clasificación y una penalización de divergencia entre predicciones limpias y adversariales, ponderada por un hiperparámetro λ:

$$\mathcal{L}_{TRADES}(f, x, y) = \mathcal{L}_{CE}(f(x), y) + \frac{1}{\lambda} \max_{\delta \in S} KL(f(x), f(x + \delta)) \tag{6.1}$$

En la práctica, λ define el compromiso entre la precisión y robustez: valores grandes priorizan la precisión, mientras que valores pequeños refuerzan la robustez adversarial. No obstante, estudios recientes como (Li et al., 2024) evidencian que TRADES presenta inestabilidad y sobrestimación de robustez cuando λ se mantiene fijo durante el entrenamiento. Diferentes configuraciones de λ producen comportamientos distintos frente a ataques, generando que el modelo sea inconsistente.

Con base en esta limitación, presentamos D-TRADES, una extensión dinámica donde el parámetro λ cambia durante el entrenamiento según las características estadísticas y de sensibilidad de los datos. En lugar de un valor constante, se define la función de $\lambda(x)$ que ajusta el grado de regularización adversarial en función de la incertidumbre de predicción y la sensibilidad adversarial de cada ejemplo:

$$\lambda(x) = \alpha \cdot \mathcal{H}(f(x)) + \beta \cdot \|\nabla_{x'} KL(f(x) \,\|\, f(x'))\|_2 \Big|_{x'=x+\delta} \tag{6.2}$$

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 $\mathcal{H}(f(x))$ es la entropía de predicción del modelo sobre el input limpio x:

$$\mathcal{H}(f(x)) = -\sum_{c=1}^{C} f(x)_c \log f(x)_c \tag{6.3}$$

Este término mide la incertidumbre epistemológica, o sea, cuando el modelo no está seguro de la clase, la entropía es alta. Por lo tanto, una entropía alta aumenta $\lambda(x)$, reforzando la regularización adversarial en ejemplos adversariales. Por otro lado, $\|\nabla_{x'} KL(f(x) \,\|\, f(x'))\|_2$ mide la sensibilidad adversarial local, es decir, qué tan rápido cambia la divergencia entre las distribuciones de salida ante pequeñas perturbaciones en la entrada; una entrada grande implica una región del espacio de entrada con alta vulnerabilidad a ataques; en consecuencia, $\lambda(x)$ crece para fortalecer la defensa contra esa muestra.

La formulación de $\lambda(x)$ se inspira en dos líneas de trabajo previas que abordan el equilibrio entre precisión y robustez desde perspectivas complementarias. Primero, Stutz et al., 2020 introducen el entrenamiento adversarial calibrado por confianza (CCAT), donde se reduce la confianza del modelo en ejemplos 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 ejemplos con alta entropía reciban mayor ponderación adversarial. Por otro lado, Ross and Doshi-Velez, 2018 proponen la regularización del gradiente de entrada, 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 ejemplo.

Una vez definida $\lambda(x)$, esta se incorpora directamente en la función de pérdida de TRADES, dando lugar a la siguiente formulación:

$$\mathcal{L}_{D\text{-}TRADES}(f, x, y) = \mathcal{L}_{CE}(f(x), y) + \lambda(x) \cdot KL(f(x) \,\|\, f(x + \delta)) \tag{6.4}$$

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 mitigando la inestabilidad reportada por Li et al., 2024, en la figura 6.1 se muestra el diagrama general del metodo propuesto, donde el modelo recibe ejemplos limpios y adversariales, calcula las salidas $f(x)$ y $f(x')$, 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 natural y la robustez durante la retropropagación, preservando la estructura teórica del TRADES original.

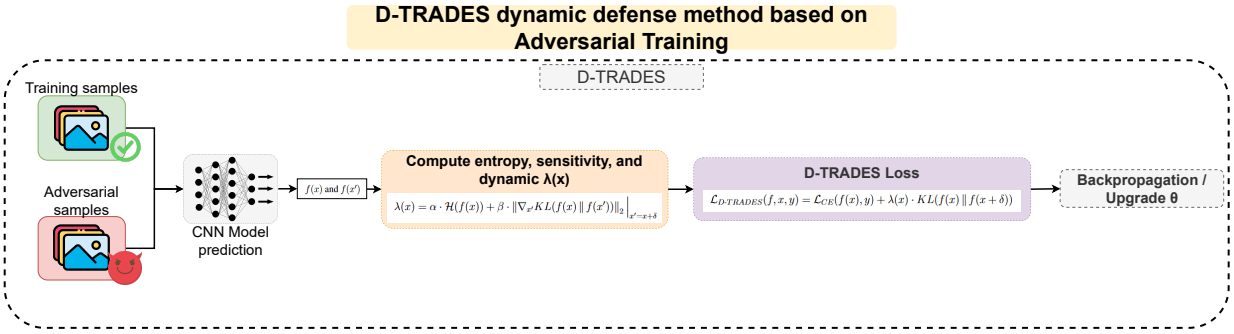

Figura 6.1: Diagrama general del método propuesto **D-TRADES**

Teóricamente, al introducir $\lambda(x)$ en la formula de TRADES, la gradiente total de la pérdida se vuelve dependiente de la entropía y la sensibilidad local, lo que induce a un efecto de autoajuste: muestras adversariales con alta incertidumbre reciben mayor valor en la función λ, mientras que muestras limpias sufren menos regularización en esa función. Este mecanismo debería reducir el sobreajuste adversarial y estabilizar la superficie de decisión, favoreciendo una robustez más homogénea a lo largo del espacio de entrada.

7. EXPERIMENTOS

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.

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	

Tabla 8.1: Planificación de actividades semestral

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

Medi, T., Jung, S., & Keuper, M. (2025). Trix- trading adversarial fairness via mixed adversarial training. https://arxiv.org/abs/2507.07768

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