Optimización evolutiva multi-objetivo de arquitecturas neuronales como estrategia de defensa ante ataques adversarios en restauración de imágenes

Trabajo Terminal No. 2024-B147

Alumnos: *Gamboa Sandoval Isabel, Ortiz Camacho Luis Enrique

Directores: Buitrón Dámaso Israel, Falcón Cardona Jesús Guillermo

*Correo electrónico: igamboas1900@alumno.ipn.mx

Resumen — Las redes neuronales profundas desempeñan un papel crucial en numerosas aplicaciones en las ciencias, ingenierías e industria, particularmente en el campo de la restauración de imágenes. Sin embargo, su función es vulnerable a ataques adversarios, lo cual plantea desafíos significativos. Considerando estos desafíos, se ha propuesto encontrar arquitecturas de redes más resistentes a ataques a través de la búsqueda de arquitecturas neuronales. En este proyecto, se diseñará y se implementará un sistema que proponga arquitecturas resistentes a ataques adversarios mediante optimización multiobjetivo, con el objetivo de reducir la vulnerabilidad de sistemas de restauración de imágenes que hagan uso de redes neuronales profundas

Palabras clave — Algoritmos Evolutivos, Arquitecturas profundas, Optimización multiobjetivo, Restauración de imágenes.

1. Introducción

En diferentes áreas de la ciencia, la ingeniería y la industria se presentan problemas donde se necesitan optimizar de forma simultánea múltiples objetivos que están en conflicto mutuo. Este tipo de problemas son los denominados *problemas de optimización multiobjetivo* (POM) [1]. En otras palabras, en este tipo de problemas se puede observar que mejorar un objetivo implica el deterioro de al menos uno de los demás objetivos. Los POMs son de gran interés puesto que modelan de una mejor manera los problemas que nacen en distintas áreas. Comprender las diferentes técnicas para resolver POMs es un tema de gran interés para científicos, ingenieros, y practicantes puesto que los POMs son usualmente muy complejos. En POMs continuos, las técnicas de programación matemática son ideales puesto que pueden garantizar la generación de soluciones óptimas. No obstante, esta garantía sólo se da bajo estrictas condiciones como diferenciabilidad de los objetivos. En consecuencia, otras técnicas se han desarrollado en los últimos años, siendo los *algoritmos evolutivos multiobjetivo* (AEMOs) unos de los prometedores [2]. Los AEMOs son técnicas estocásticas poblacionales que no requieren información de la derivada, lo cual les permite resolver POMs cuyos objetivos no están definidos de forma analítica (por ejemplo, cuando los valores objetivo provienen de simulaciones) o son altamente no lineales. Es por esto que los AEMOs han tenido éxito en la resolución de POMs altamente complejos donde las técnicas de programación matemática no han generado buenos resultados.

La *inteligencia artificial* (IA) busca simular comportamientos inteligentes, a menudo con el objetivo de que las máquinas despachen una gran cantidad de información. Dentro de esta disciplina se encuentra el aprendizaje profundo, el cual estudia diferentes arquitecturas neuronales multicapa diseñadas para tareas de clasificación o regresión, entre otras, basadas en grandes volúmenes de datos. En el aprendizaje profundo se pueden encontrar las redes neuronales profundas (*Deep Neural Networks* o DNNs, por sus siglas en inglés) y las redes neuronales convolucionales (*Convolutional Neural Network* o CNNs, por sus siglas en inglés) [3]. Las DNNs, son modelos que constan de múltiples capas de neuronas interconectadas, lo que les permite aprender representaciones complejas a partir de datos brutos. Estas redes son particularmente adecuadas para tareas que requieren un alto grado de abstracción, como el procesamiento de lenguaje natural y el reconocimiento de voz [4]. Por otro lado, las CNNs, que son una variante de las DNNs, han sido diseñadas específicamente para tareas relacionadas con la visión por computadora. Estas redes son altamente eficientes en la extracción de características visuales de

imágenes, por lo que son muy usadas en los campos de reconocimiento de objetos, la detección de rostros y la clasificación de imágenes [5].

Un desafío actual para el aprendizaje profundo son los ataques adversarios. En el caso de procesamiento de imágenes, este tipo de ataques consiste en que un adversario añada una perturbación a una imagen que va a ser presentada a una red neuronal con el objeto de que el modelo la clasifique correctamente, produciendo un bajo rendimiento del sistema de aprendizaje [6]. Varios estudios han demostrado la efectividad de ataques basados en el cálculo de gradientes para engañar a DNNs que trabajan con imágenes [7], [8], [9], [10], mostrando que hasta la alteración de un único píxel afecta los resultados del modelo atacado [7], [9]. Las modificaciones en los ejemplos adversarios suelen ser imperceptibles para el ojo humano, presentando un riesgo significativo para los sistemas que usen DNNs. Es por esto que es de suma importancia el diseño de sistemas de protección ante los ataques adversarios. Con el objetivo de entender mejor las arquitecturas neuronales y cómo defenderlas ante ataques adversarios, investigamos el estado del arte de diferentes tipos de ataques adversarios (véase Tabla 1) y estrategias de defensa (véase Tabla 2):

La restauración de imágenes es un campo del procesamiento de imágenes que se ha vuelto de gran interés, en los últimos años debido a su amplia gama de aplicaciones . En general, un método de restauración de imágenes parte de una imagen alterada (por ejemplo, con difuminación o baja resolución, ruido o desenfoque) y su meta es obtener una versión mejorada y más fiel ésta, es decir, filtrar las imperfecciones. Para lograr esto, se recurre a una variedad de técnicas y algoritmos, incluyendo enfoques basados en inteligencia artificial, como las DNNs que han demostrado ser altamente efectivas en esta tarea [11], [12], [13], [14].

En el contexto de los desafíos anteriormente descritos, nuestro objetivo es diseñar e implementar un sistema de búsqueda de arquitecturas neuronales que proponga arquitecturas más resistentes a ataques adversarios mediante AEMOs. Esto servirá como una estrategia de defensa efectiva contra ataques adversarios que pudieran comprometer la integridad y rendimiento de sistemas que utilicen DNNs para la restauración de imágenes.

2. Planteamiento del problema

La aplicación de las DNNs en la restauración de imágenes ha revolucionado diversos campos, por ejemplo, la medicina, la fotografía, la vigilancia y la visión por computadora [3], [15]. Sin embargo, este avance ha suscitado una creciente preocupación debido a la vulnerabilidad de estos sistemas ante ataques adversarios [8]. Estos ataques pueden socavar seriamente el desempeño y la confiabilidad de las redes neuronales profundas utilizadas en la restauración de imágenes [16].

El problema central de este trabajo terminal consiste en el diseño e implementación de un sistema automatizado que mejore arquitecturas ya existentes. Para este fin, se plantea la búsqueda de arquitecturas neuronales, como una vía alternativa, para resistir ante ataques adversarios. En consecuencia, la búsqueda de arquitecturas neuronales se modela como un POM que puede ser resuelto mediante un AEMO.

La justificación para emplear AEMOs en este problema radica en su capacidad para alcanzar simultáneamente diversos objetivos, más allá de la simple reducción del tiempo de búsqueda o el error de generalización, como por ejemplo el uso de memoria o energía, la latencia de salida, o en nuestro caso, la robustez contra ataques adversarios. De esta forma, se permite una exploración más ágil del espacio de búsqueda, al permitir flexibilidad al ajustar los compromisos entre los diferentes objetivos a alcanzar. En particular, los algoritmos evolutivos destacan por operar sobre una población de soluciones candidatas, lo que facilita la exploración paralela del espacio de búsqueda. Esto se traduce en una búsqueda más inteligente comparada con métodos de optimización de un solo punto [44].

La búsqueda de arquitecturas neuronales es un problema muy complejo del estado del arte, la cual involucra espacios de búsquedas complejos y de múltiples dimensiones, no lineales, discontinuos e incluso en ocasiones presentan una combinación de ambas. A través de las operaciones de mutación y recombinación, los algoritmos evolutivos pueden navegar a través de estos espacios para encontrar soluciones aproximadamente óptimas de una manera inteligente. La naturaleza no lineal y no continua de las decisiones arquitecturales, derivada de los

hiperparámetros de la arquitectura, impide la aplicación directa de técnicas basadas en la optimización por gradiente. Por lo tanto, es necesario recurrir a algoritmos evolutivos que puedan manejar la naturaleza discreta y discontinua del espacio de búsqueda [45].

Desde la perspectiva de los *algoritmos evolutivos* (AEs) como técnica para NAS, la representación de las DNNs se convierte en un desafío importante debido a su alta dimensionalidad y complejidad. La representación cromosómica utilizada en algoritmos genéticos para optimización de un solo objetivo, resulta inadecuada. Por tanto, se hace necesaria la exploración y desarrollo de representaciones alternativas que sean más adecuadas para la optimización de DNNs [34], [35]. Otro desafío crítico radica en el proceso de entrenamiento de las poblaciones candidatas. Esto es esencial para evaluar el desempeño de las arquitecturas propuestas, pero representa un desafío computacional significativo debido a su alto costo en tiempo y recursos. Así mismo, se hace necesario diseñar un sistema de toma de decisiones que permita seleccionar las soluciones que aproximen de mejor forma la solución a un POM, proporcionando una estrategia sólida para la implementación práctica de las arquitecturas encontradas en entornos reales.

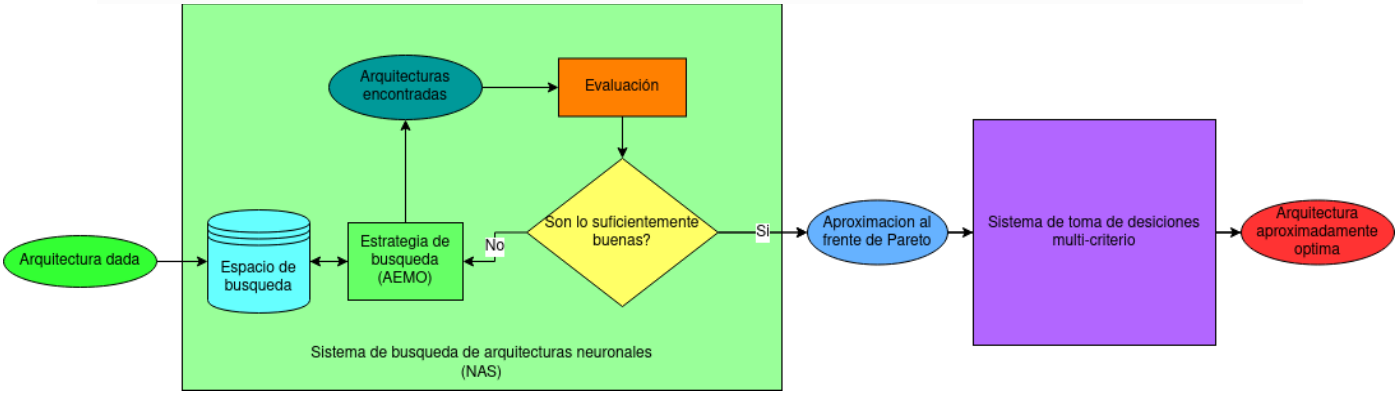

Figura 1. Arquitectura del proyecto

Tabla 1. Revisión del estado del arte sobre ataques adversarios a DNNs y CNNs.

Recopilación de ataques adversarios del estado del arte				
Título	**Tipo de Ataque**	**¿Cómo se genera el ataque?**	**¿Qué tipo de redes neuronales atacan?**	**¿Qué tipos de datos ataca?**
One-pixel and X-pixel adversarial attacks based on smell bees optimization algorithm	One pixel attack X-Pixel Attack Non-targeted attack	Se escoge un conjunto de píxeles al azar y se minimizan con el algoritmo SBO [7].	CNN	Imágenes
Point Cloud Adversarial Perturbation Generation for Adversarial Attacks	Adversarial attack Point Cloud attack	Añade nuevos puntos a una nube para generar perturbaciones [19].	CNN	Modelos 3D
DCVAE-adv*: A Universal Adversarial Example Generation Method for White and Black Box Attacks*	White-box attack Black-box attack	Se entrena una red neuronal para atacar a otra, se penalizan los ejemplos adversarios más inconsistentes [8].	CNN	Clases, imágenes
The race to robustness: exploiting fragile models for urban camouflage and the imperative for machine learning security	Universal adversarial perturbation FGSM PGD One Pixel Attack Adversarial Patch Targeted attack	Busca modificar las regiones más pequeñas que tengan mayor impacto en la clasificación [9].	CNN	Imágenes
Understanding Adversarial Attacks on Deep Learning BasedMedical Image Analysis Systems	Adversarial attack	Se hace una versión iterativa del ataque FSGM [10].	CNN	Imágenes médicas
One Pixel Adversarial Attacks via Sketched Programs	Adversarial attack	Se tiene una estructura predefinida para identificar ejemplos adversos [20].	CNN	Imágenes
ZOO: *Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models*	Adversarial attack	Explotar el zeroth order optimization para estimar directamente los gradientes de la DNN objetivo para generar ejemplos adversarios [21].	DNN	Imágenes

Few pixels attacks with generative model	Adversarial attack	Una red convolucional con un auto-encodificador variacional es entrenado. Se utiliza la pérdida adversaria y la pérdida de reconstrucción para aprender las muestras adversarias [22].	CNN	Imágenes
PISA: Pixel skipping-based attentional black-box adversarial attack	Adversarial attack	PISA utiliza la técnica de mapeo de activación de clase (CAM) para identificar píxeles en la región saliente. Luego los refina para reducir su número [23].	DNN	Imágenes
DeepFool: a simple and accurate method to fool deep neural networks	Adversarial attack	Encontrar la perturbación mínima para cambiar la clasificación de una imagen mediante el cálculo del gradiente y su repetición hasta lograr la modificación [24].	CNN	Imágenes
AESOP: Adjustable Exhaustive Search for One-Pixel Attacks in Deep Neural Networks	One-pixel attack	Algoritmo de búsqueda exhaustiva ajustable que evalúa el impacto de agregar un píxel en la imagen, en función de las características de la imagen de entrada y la red neuronal atacada [25].	DNN	Imágenes
SparseFool: a few pixels make a big difference	Adversarial attack	Explota la baja curvatura media del límite de decisión. El algoritmo se aproxima iterativamente al límite de decisión y actualiza la imagen perturbadora hasta que cambia la etiqueta de la red [26].	CNN	Imágenes

Tabla 2. Revisión del estado del arte sobre estrategias de defensa ante ataques adversarios a DNNs y CNNs.

Recopilación de métodos de defensa ante ataque adversarios del estado del arte					
Título	**¿Ante qué tipo de ataques defiende?**	**Estrategia de defensa**	**¿Qué tipo de datos son atacados?**	**Razón de la defensa**	**¿Qué redes neuronales defiende?**
NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data	Ataques adversariales	Se modifican los grafos de manera que sean difíciles de atacar sin reducir su rendimiento [15].	Graph data Datos privados	Proteger la privacidad de los datos.	GNN
A Robust CycleGAN-L2 Defense Method for Speaker Recognition System	White-box attacks	Se entrena a una red neuronal con ejemplos adversarios y muestras originales para que sea capaz de identificar la fuente [27].	Audio	Proteger al modelo original con el mínimo impacto en su precisión.	CNN
Encryption inspired adversarial defence for visual classification	Adversarial attacks White-box attacks	El modelo puede detectar ataques con base en la llave introducida para decodificar los datos [28].	Imágenes RGB de 8-bits CIFAR-10	La mayoría de las defensas reducen la precisión del modelo y de cualquier forma estas suelen ser derrotadas.	DNN
An Online Website Fingerprinting Defense Based on the Non-Targeted Adversarial Patch	Website Fingerprinting	Se utiliza el algoritmo Grad-CAM para determinar los segmentos críticos y se inyecta un parche para defender la red [29].	Traffic trace Paquetes	Implementar una defensa que pueda proteger datos en línea.	CNN
Detection Mechanisms of One-Pixel Attack	One Pixel Attack	Busca un píxel que podría causar una mala clasificación [30].	CIFAR-10	Los métodos propuestos son los primeros en enfocarse en ataques de un píxel.	CNN
The Best Defense is a Good Offense: Adversarial Augmentation Against Adversarial Attacks	White-box attack Grey-box attack Black-box attack MitM attacks	Combina diferentes estrategias de defensa [31].	Imágenes MNIST CIFAR10 FashionMNIST TinyImageenet	La filosofía detrás de los distintos tipos de defensa pueden ser adaptadas entre ellas.	DNN
Analysis of the Effect of Adversarial Training in Defending EfficientNet-B0 Model from DeepFool Attack	Adversarial attack	Se usan imágenes adversarias para que el modelo se adapte a este tipo de ataque [32].	Imágenes RGB de 224*224 pixeles Malarial Cell Image Dataset Lung CT Scan image dataset	Mejorar la robustez de los modelos usados para realizar diagnósticos médicos.	CNN EfficientNet-B0

3. Justificación

La creación de sistemas automáticos de búsqueda de arquitecturas neuronales se ha vuelto una necesidad inminente, impulsada por las crecientes complejidades que rodean a las DNNs. Esta complejidad nace de múltiples factores fundamentales que tornan ineficiente y poco práctico el proceso manual de diseño de arquitecturas neuronales [36].

En primer lugar, las DNN modernas suelen contar con una abrumadora cantidad de parámetros [3], a menudo en el rango de millones. La definición manual de estas arquitecturas se torna propensa a errores humanos y altamente ineficiente. Regularmente, solo los diseñadores de una red neuronal en específico conocen, en alguna medida, su comportamiento, por lo que son los más indicados para establecer alguna arquitectura neuronal. No obstante, sin una caracterización matemática de la arquitectura de una red neuronal, la tarea de diseño se tiñe de subjetividad. En consecuencia, es necesario contar con mecanismos algorítmicos capaces de construir de forma automática la arquitectura de una DNN para cualquier tipo de problema sin la intervención humana. Esto facilitaría el uso de las DNNs por personas diferentes a los diseñadores, ampliando el acceso a este tipo de tecnología.

Además, el diseño de una arquitectura neuronal se encuentra intrínsecamente vinculado al problema que se pretende resolver. Diferentes tareas, como la clasificación de imágenes, el procesamiento de lenguaje natural o la detección de anomalías, demandan estructuras de red distintas [15], [27], [28], [31]. Incluso dentro de una misma tarea, los requisitos pueden variar drásticamente según el conjunto de datos y las restricciones específicas del problema.

Esta complejidad se magnifica ante la amenaza constante de ataques adversarios en sistemas de información. Las DNN son conocidas por su susceptibilidad a tales ataques [6], lo que resulta inaceptable en aplicaciones críticas, como el diagnóstico médico [24], la seguridad en vehículos autónomos [37] y sistemas financieros [38]. La evolución constante de nuevas estrategias para evadir sistemas de seguridad agrava aún más este desafío. Por consiguiente, se hace imperativo la búsqueda de arquitecturas neuronales resistentes ante ataques adversarios para desarrollar soluciones sólidas en aplicaciones críticas.

Los sistemas automáticos de búsqueda de arquitecturas neuronales ofrecen una solución clave a estos desafíos. Una solución viable es la utilización de técnicas de optimización, como los AEMOs [2], dado el NAS como un POM. Los AEMOs tienen la capacidad de explorar heurísticamente un vasto espacio de búsqueda con el objeto de encontrar arquitecturas aproximadamente óptimas para tareas específicas. Es por esto que su empleo es potencial como mecanismo de defensa ante ataques adversarios. Lo que distingue a los AEMOs es su capacidad para explorar y optimizar múltiples objetivos de manera simultánea. En el contexto de la búsqueda automática de arquitecturas neuronales, esto implica que los AEMOs pueden generar una amplia variedad de arquitecturas neuronales, cada una de las cuales ofrece diferentes compensaciones entre objetivos. Estos objetivos pueden incluir precisión de clasificación, complejidad del modelo, robustez, tiempo de entrenamiento, entre otros. Esta diversidad intrínseca fortalece la resistencia de la red al evitar el sobreajuste a tipos específicos de ataques adversarios [32], dificultando que los atacantes exploten vulnerabilidades comunes a todas las arquitecturas.

La aplicación de sistemas automáticos de búsqueda de arquitecturas neuronales se extiende a diversos campos, desde la atención médica hasta la ciberseguridad y la inteligencia artificial [3], brindando una solución versátil para abordar los desafíos de seguridad. En el caso de aplicaciones de restauración de imágenes, su impacto es igualmente significativo. Estas aplicaciones incluyen campos como la medicina, donde se emplean en filtrado de ruido de radiografías, la mejora de imágenes de resonancia magnética y la reconstrucción de imágenes de tomografía computarizada [39]. Garantizar la seguridad de las DNN a través de los AEMOs podría asegurar la precisión y confiabilidad de los diagnósticos médicos, al mismo tiempo que protege los datos sensibles de los pacientes de manipulaciones adversarias.

En las imágenes satelitales y capturadas por drones, la restauración de imágenes realiza aportes significativos en la clasificación de especies animales [40], el monitoreo ambiental y la lucha contra el cambio climático [41] y desastres naturales [42]. Al salvaguardar las DNNs de la interferencia adversaria, se preserva la integridad de los datos utilizados, por ejemplo, en la toma de decisiones por organizaciones civiles y/o gubernamentales.

En sistemas de videovigilancia, asegurar estas redes neuronales mantiene la confiabilidad de las grabaciones de seguridad, las cuales son esenciales en la prevención del crimen y la investigación de incidentes [43]. Estas aplicaciones, junto con muchas otras, dependen de sistemas automáticos de búsqueda de arquitecturas neuronales optimizadas a través de AEMOs para proteger su integridad y funcionamiento en un mundo donde la ciberseguridad y la precisión son críticas.

Este proyecto no solo aporta originalidad al emplear AEMOs como enfoque de defensa contra ataques adversarios, sino que también mejora significativamente la resistencia de sistemas de restauración de imágenes en aplicaciones diversas y críticas.

4. Objetivo

El objetivo general de este proyecto terminal se enuncia a continuación: Buscar arquitecturas neuronales para restauración de imágenes que sean resistentes ante ataques adversarios a través de la implementación de un AEMO con el objeto de reducir la vulnerabilidad de estos sistemas que hacen uso de DNNs. Para ello es necesario definir la búsqueda de arquitecturas neuronales como un POM. Además, se evaluará el desempeño del algoritmo propuesto a través de problemas de prueba estándar y medidas calidad del estado del arte.

Los objetivos específicos de este trabajo terminal que coadyuvará en el cumplimiento del objetivo general son los siguientes:

1. Realizar un estudio exhaustivo sobre DNNs aplicadas en restauración de imágenes para conocer sus arquitecturas propuestas e identificar posibles mejoras que se puedan obtener a través de un sistema de NAS.
2. Realizar un estudio de las propuestas actuales sobre NAS planteadas como un POM con el fin de conocer sus ventajas y desventajas.
3. Plantear un esquema de codificación para DNNs de tal suerte que esta codificación sea explotada por un AEMO de forma más eficiente, aumentando la probabilidad de buscar en regiones prometedoras del espacio de búsqueda.
4. Realizar un estudio exhaustivo sobre ataques adversarios para DNNs que manejan restauración de imágenes con el objeto de conocer sus vulnerabilidades y así colocar mayor atención en la búsqueda de arquitecturas que alivien estos problemas.
5. Probar experimentalmente el desempeño del AEMO propuesto a través de problemas de prueba estándar y medidas de calidad del estado del arte con el objeto de comparar su desempeño con respecto a propuestas del estado del arte.
6. Implementar un sistema de toma de decisiones multicriterio que selecciona una de las soluciones que conforma la aproximación al conjunto de Pareto que genera el AEMO para implementar esta solución en la práctica y medir su desempeño ante ataques adversarios.

5. Productos o resultados esperados

A partir del desarrollo de este trabajo terminal se esperan los siguiente productos:

1. Un esquema de codificación de DNNs que sea explotado por un AE.
2. Un AEMO que implemente una NAS como mecanismo de defensa ante ataques adversarios en sistemas de restauración de imágenes.
3. Un sistema de soporte de toma de decisiones para seleccionar una arquitectura neuronal del conjunto de aproximación al conjunto de Pareto.
4. La escritura de un artículo científico para su publicación en algún foro nacional o internacional.

Los resultados esperados de este proyecto son los siguientes:

1. Se espera que el desempeño del AEMO sea al menos competitivo con respecto a lo reportado en el estado del arte.
2. Se espera que la red neuronal encontrada sea resistente ante diferentes ataques adversarios de tal manera que se tenga cierto umbral de confiabilidad para operar en un ambiente con condiciones similares a la reales.

5. Metodología

La metodología de este trabajo terminal está basada en cuatro etapas (véase Figura 1) de trabajo que respetan los lineamientos del método científico. A continuación, se describen estas etapas del desarrollo metodológico y se describen las actividades englobadas.

1. **Etapa 1**. El objetivo de esta fase es la investigación exhaustiva de métodos del estado del arte relacionados con las áreas de trabajo del proyecto terminal. Estas áreas comprenden: redes neuronales profundas para tareas de restauración de imágenes, algoritmos evolutivos multi-objetivo para búsqueda de arquitecturas neuronales, y ataques adversarios a redes neuronales que procesan imágenes.
    a. **Actividad 1.** Realizar un estudio exhaustivo sobre redes neuronales profundas del estado del arte para procesamiento de imágenes. Hacer especial énfasis en aplicaciones de restauración de imágenes.
    b. **Actividad 2.** Realizar un estudio exhaustivo sobre algoritmos evolutivos multiobjetivo del estado del arte aplicados a problemas de búsqueda de arquitecturas neuronales profundas. Identificar los tipos de codificación de las arquitecturas neuronales.
    c. **Actividad 3**. Realizar un estudio exhaustivo sobre ataques adversarios a redes neuronales profundas y los métodos de defensa. Identificar las vulnerabilidades más comunes y las menos comunes de estos sistemas neuronales.
    d. **Actividad 4**. Estudiar las técnicas de toma de decisiones multi-criterio del estado del arte. Hacer énfasis en investigar aquellos mecanismos que hayan sido empleados para la toma de decisiones sobre redes neuronales profundas.
2. **Etapa 2**. La meta de esta etapa se centra en realizar pruebas de los métodos estudiados en la fase anterior. Es decir, se deben reproducir los resultados de las propuestas más significativas que se hayan encontrado en el estado del arte de tal arte que sirvan como base para la construcción de la propuesta algorítmica.
    a. **Actividad 5**. Probar las redes neuronales profundas más importantes del estado del arte para procesamiento de imagen. Se deberán emplear conjuntos de problemas de prueba estándar y medidas de desempeño del estado del arte. Se deben comparar los resultados obtenidos con aquellos reportados en el estado del arte.
    b. **Actividad 6**. Probar los algoritmos evolutivos del estado del arte para búsqueda de arquitecturas neuronales. Es necesario medir el tiempo de cómputo de estos algoritmos y determinar si es común el uso de modelos subrogados. Se debe hacer énfasis en estudiar la eficiencia de los mecanismos de representación de redes neuronales para un algoritmo evolutivo.
    c. **Actividad 7.** Probar los ataques más comunes a redes neuronales profundas para procesamiento de imágenes. Determinar las vulnerabilidades que atacan y medir la degradación de desempeño de las redes neuronales profundas.
    d. **Actividad 8**. Probar los mecanismos de defensa ante ataques adversarios sobre redes neuronales profundas que procesan imágenes. Medir la ganancia en desempeño que se obtiene al implementar cada estrategia de defensa.
3. **Etapa 3.** Esta fase del desarrollo se centra en el establecimiento de la búsqueda de arquitecturas neuronales como un problema de optimización multi-objetivo. En consecuencia, también se diseñará un algoritmo evolutivo multi-objetivo que implemente las fortalezas estudiadas y experimentadas en las etapas anteriores (Actividades 2 y 5, respectivamente). Finalmente, se debe integrar un sistema de toma de decisiones multi-criterio para seleccionar una arquitectura encontrada.

a. **Actividad 9**. Definir la búsqueda de arquitecturas neuronales como un problema de optimización multi-objetivo.
b. **Actividad 10**. Diseñar un algoritmo evolutivo multi-objetivo que busque arquitecturas neuronales aproximadamente óptimas que resistan ante los ataques adversarios más comunes que se presentan en la literatura.
c. **Actividad 11**. Medir el desempeño del algoritmo propuesto mediante medidas de calidad del estado del arte y compararlos con respecto a otros algoritmos del estado del arte.
d. **Actividad 12**. Estudiar las arquitecturas encontradas y caracterizar las propiedades de un arquitectura potencial a ser implementada y que cumpla con criterios de un tomador de decisión.
e. **Actividad 13**. Implementar un mecanismo de toma de decisiones multi-criterio que explore las soluciones encontradas mediante el algoritmo evolutivo multi-objetivo. Probar la solución seleccionada ante ataques adversarios para los cuáles fue construida y para aquellos para los que no.

4. **Etapa 4**. Esta fase se centra en la escritura de resultados del proyecto terminal así como contemplar la publicación de algún artículo científico en algún foro con arbitraje.
a. **Actividad 14**. Escritura del reporte para Proyecto Terminal 1.
b. **Actividad 15**. Escritura del reporte para Proyecto Terminal 2.
c. **Actividad 16 (opcional)**. Escritura de un artículo científico para algún foro nacional o internacional con arbitraje estricto.

Figura 1. Etapas del proyecto.

6. Cronograma

A continuación se muestra una propuesta del cronograma de actividades (véase la Figura 2) con base en las tareas declaradas en la metodología de desarrollo del proyecto terminal Cabe destacar que cada actividad está asociada con algún estudiante en particular o asignada a ambos estudiantes.

Actividad	Responsable	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Actividad 1	**Enrique**											
Actividad 2	**Isabel**											
Actividad 3	**Enrique**											
Actividad 4	**Isabel**											
Actividad 5	**Enrique**											

Actividad 6	**Isabel**											
Actividad 7	**Ambos**											
Actividad 8	**Ambos**											
Actividad 14	**Ambos**											
Actividad 9	**Ambos**											
Actividad 10	**Ambos**											
Actividad 11	**Isabel**											
Actividad 12	**Enrique**											
Actividad 13	**Ambos**											
Actividad 15	**Ambos**											
Actividad 16	**Ambos**											

Figura 2. Cronograma de actividades.

7. Referencias

[1] M. Emmerich y A. H. Deutz, "A tutorial on multiobjective optimization: fundamentals and evolutionary methods", Natural Computing, Vol. 17, Num. 3, Pags. 585-609, mayo de 2018, DOI: 10.1007/s11047-018-9685-y.

[2] K. Deb y K. Deb, Multi-Objective optimization using evolutionary algorithms. 2001. [En línea]. Disponible en: http://ci.nii.ac.jp/ncid/BB00925127.

[3] I. H. Sarker, "Deep Learning: a comprehensive overview on techniques, taxonomy, applications and research directions", SN computer science, Vol. 2, Num. 6, agosto de 2021, DOI: 10.1007/s42979-021-00815-1.

[4] Y. Goldberg, “A Primer on Neural Network Models for Natural Language Processing”, Journal of Artificial Intelligence Research, Vol. 57. AI Access Foundation, Pags. 345–420, noviembre de 2016. DOI: 10.1613/jair.4992.

[5] D. C. Ciresan et al., "Flexible, High Performance Convolutional Neural Networks for Image Classification," International Joint Conference on Artificial Intelligence, 2011.

[6] M. Ozdag, "Adversarial attacks and Defences against deep neural networks: a survey", Procedia Computer Science, Vol. 140, Pags. 152-161, enero de 2018, DOI: 10.1016/j.procs.2018.10.315.

[7] Y. M. B. Ali, "One-pixel and X-pixel adversarial attacks based on smell bees optimization algorithm," *Future Generation Computer Systems*, Vol. 149, Pags. 562-576, 2023.

[8] L. Xu y J. Zhai, "DCVAE-adv: A Universal Adversarial Example Generation Method for White and Black Box Attacks", Tsinghua Science & Technology, Vol. 29, Num. 2, Pags. 430-446, abril de 2024, DOI: 10.26599/tst.2023.9010004.

[9] H. Farlow, M. A. Garratt, G. Mount, y T. Lynar, "The race to robustness: exploiting fragile models for urban camouflage and the imperative for machine learning security", arXiv (Cornell University), junio de 2023, DOI: 10.48550/arxiv.2306.14609.

[10] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, y Feng Lu. "Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems." *arXiv preprint arXiv:2002.05638* (2020).

[11] C. Mou, Q. Wang, y J. Zhang, “Deep Generalized Unfolding Networks for Image Restoration”. arXiv, 2022. DOI: 10.48550/ARXIV.2204.13348.

[12] W. Dong, P. Wang, W. Yin, G. Shi, F. Wu, y X. Lu, “Denoising Prior Driven Deep Neural Network for Image Restoration”, arXiv, 2018.

[13] Z. Wang, J. Chen, S. C. H. Hoi, Deep learning for image super-resolution: A survey (2020). arXiv:1902.06068.

[14] H. Zhang, Y. Dai, H. Li, P. Koniusz, Deep stacked hierarchical multi-patch network for image deblurring, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[15] I.-C. Hsieh y C.-T. Li, "NetFense: Adversarial Defenses against Privacy Attacks on neural networks for graph data", IEEE Transactions on Knowledge and Data Engineering, Pag. 1, enero de 2022, DOI: 10.1109/tkde.2021.3087515.

[16] Ian J. Goodfellow, Jonathon Shlens and Christian Szegedy, "Explaining and harnessing adversarial examples", CoRR abs/1412.6572, 2015.

[17] S. Ali y M. A. Wani, “Gradient-Based Neural Architecture Search: A Comprehensive Evaluation”, Machine Learning and Knowledge Extraction, Vol. 5, Num. 3. MDPI AG, Pags. 1176–1194, septiembre de 2014, 2023. DOI: 10.3390/make5030060.

[18] K. Kandasamy, W. Neiswanger, J. Schneider, B. Poczos, y E. Xing, “Neural Architecture Search with Bayesian Optimisation and Optimal Transport”, arXiv, 2018.

[19] F. He, Y. Chen, R. Chen, y W. Nie, "Point cloud adversarial perturbation generation for adversarial attacks", IEEE Access, Vol. 11, Pags. 2767-2774, enero de 2023, DOI: 10.1109/access.2023.3234313.

[20] T. Yuviler y D. Drachsler-Cohen, "One Pixel Adversarial Attacks via Sketched Programs," Proc. ACM Program. Lang., Vol. 7, PLDI, Art. 187, 2023.

[21] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, y C.-J. Hsieh, "ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models”, Proceedings of the ACM Workshop on Artificial Intelligence and Security (AISec), noviembre de 2017

[22] Y. Li, Q. Pan, Z. Feng y E. Cambria, “Few pixels attacks with generative model”, *Pattern Recognit.*, Pag. 109849, julio de 2023.

[23] J. Wang, Z. Yin, J. Jiang, J. Tang y B. Luo, “PISA: Pixel Skipping-based Attentional Black-box Adversarial Attack”, *Comput. & Secur.*, Pag. 102947, octubre de 2022.

[24] S. -M. Moosavi-Dezfooli, A. Fawzi y P. Frossard, "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016

[25] W. Nam y H. Kil, “AESOP: Adjustable Exhaustive Search for One-Pixel Attacks in Deep Neural Networks”, *Appl. Sci.*, Vol. 13, Num. 8, Pag. 5092, abril de 2023.

[26] A. Modas, S. -M. Moosavi-Dezfooli y P. Frossard, "SparseFool: A Few Pixels Make a Big Difference," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] L. Yang, X. Yang, S. Zhang, y X. Zhang, "A robust CycleGAN-L2 defense method for speaker recognition system", IEEE Access, Vol. 11, Pags. 82771-82783, enero de 2023, DOI: 10.1109/access.2023.3300031.

[28] M. Maung, A. Pyone and H. Kiya, "Encryption Inspired Adversarial Defense For Visual Classification," 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2020, Pags. 1681-1685, DOI: 10.1109/ICIP40778.2020.9190904.

[29] X. Gu, B. Song, W. Lan and M. Yang, "An Online Website Fingerprinting Defense Based on the Non-Targeted Adversarial Patch," in Tsinghua Science and Technology, Vol. 28, no. 6, Pags. 1148-1159, diciembre de 2023, DOI: 10.26599/TST.2023.9010062.

[30] P. Wang, Z. Cai, D. Kim y W. Li, "Detection Mechanisms of One-Pixel Attack", *Wireless Commun. Mobile Comput.*, Vol. 2021, febrero de 2021.

[31] I. Frosio and J. Kautz, "The Best Defense is a Good Offence: Adversarial Augmentation Against Adversarial Attacks," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, Pags. 4067-4076, DOI: 10.1109/CVPR52729.2023.00396.

[32] A. M. A., B. M., S. D. Dunston and M. A. R. V., "Analysis of the Effect of Adversarial Training in Defending EfficientNet-B0 Model from DeepFool Attack," 2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, 2023, Pags. 1-7, DOI: 10.1109/ICCT56969.2023.10075774.

[33] Y. Liu and J. Lu, "Double Loss Block Neural Architecture Search," 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2022, Pags. 731-734, DOI: 10.1109/ITAIC54216.2022.9836540.

[34] Y. Sun, B. Xue, M. Zhang, y G. G. Yen, "Evolving Deep Convolutional Neural Networks for Image Classification", 2017.

[35] M. Suganuma, S. Shirakawa, y T. Nagao, "A Genetic Programming Approach to Designing Convolutional Neural Network Architectures". arXiv, 2017. DOI: 10.48550/ARXIV.1704.00764.

[36] R. Shi, J. Luo and Q. Liu, "Fast Evolutionary Neural Architecture Search Based on Bayesian Surrogate Model," 2021 IEEE Congress on Evolutionary Computation (CEC), Kraków, Poland, 2021, Pags. 1217-1224, DOI: 10.1109/CEC45853.2021.9504999.

[37] J. Zhang, Y. Lou, J. Wang, K. Wu, K. Lu, y X. Jia, "Evaluating Adversarial Attacks on Driving Safety in Vision-Based Autonomous Vehicles", IEEE Internet of Things Journal, Vol. 9, Num. 5. Institute of Electrical and Electronics Engineers (IEEE), Pags. 3443–3456, 01 de marzo de 2022. DOI: 10.1109/jiot.2021.3099164.

[38] I. Fursov et al., "Adversarial Attacks on Deep Models for Financial Transaction Records". arXiv, 2021. DOI: 10.48550/ARXIV.2106.08361.

[39] E. Ahishakiye, M. B. Van Gijzen, J. Tumwiine, R. Wario, y J. Obungoloch, "A survey on deep learning in medical image reconstruction", Intelligent Medicine, Vol. 1, Num. 3. Elsevier BV, Pags. 118–127, septiembre de 2021. DOI: 10.1016/j.imed.2021.03.003.

[40] V. Lopez-Vazquez, J. M. Lopez-Guede, D. Chatzievangelou, y J. Aguzzi, "Deep learning based deep-sea automatic image enhancement and animal species classification", Journal of Big Data, Vol. 10, Num. 1. Springer Science and Business Media LLC, marzo de 2023.

[41] Y. Zhu et al., "Learning Weather-General and Weather-Specific Features for Image Restoration Under Multiple Adverse Weather Conditions", 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, junio de 2023. DOI: 10.1109/cvpr52729.2023.02083.

[42] S. Karavarsamis, I. Gkika, V. Gkitsas, K. Konstantoudakis, y D. Zarpalas, "A Survey of Deep Learning-Based Image Restoration Methods for Enhancing Situational Awareness at Disaster Sites: The Cases of Rain, Snow and Haze", Sensors, vol. 22, núm. 13. MDPI AG, Pag. 4707, junio de 2022.

[43] Md. S. H. Onim et al., “BLPnet: A new DNN model and Bengali OCR engine for Automatic Licence Plate Recognition”, Array, Vol. 15. Elsevier BV, Pag. 100244, septiembre de 2022. DOI: 10.1016/j.array.2022.100244.

[44] T. Elsken, J. H. Metzen, y F. Hutter, “Neural Architecture Search: A Survey”, arXiv, 2018, doi: 10.48550/ARXIV.1808.05377.

[45] M. Wistuba, A. Rawat, y T. Pedapati, “A Survey on Neural Architecture Search”. arXiv, 2019. doi: 10.48550/ARXIV.1905.01392.

8. Alumnos y Directores

Isabel Gamboa Sandoval .- Estudiante de la licenciatura en Ingeniería en Sistemas Computacionales en ESCOM-IPN, Especialidad Sistemas, Boleta: 2020630619, Tel. +52 55-8368-4255, Correo electrónico: igamboas1900@alumno.ipn.mx.

CARÁCTER: Confidencial
FUNDAMENTO LEGAL: Artículo 11 Fracc. V y Artículos 108, 113 y 117 de la Ley Federal de Transparencia y Acceso a la Información Pública.
PARTES CONFIDENCIALES: Número de boleta y teléfono.

Firma: Isabel Gamboa

Luis Enrique Ortiz Camacho .- Estudiante de la licenciatura en Ingeniería en Sistemas Computacionales en ESCOM-IPN, Boleta: 2020630629, Tel. +52 744-269-7148, Correo electrónico: lortizc1901@alumno.ipn.mx.

Firma:

Israel Buitrón Dámaso .- Profesor de tiempo completo en la ESCOM-IPN. Es Ingeniero en Sistemas Computacionales por la ESCOM-IPN. Es Maestro en Ciencias en Computación, en el área de criptografía, por el CINVESTAV-IPN. Realizó estudios doctorales, en el área de teoría de grafos y protocolos de autenticación, en el Departamento de Computación del CINVESTAV-IPN. Teléfono: +52 55-5729-6000; 54041, Página web: comunidad.escom.ipn.mx/ibuitron. Correo electrónico: ibuitron@ipn.mx.

Firma:

Jesús Guillermo Falcón Cardona.- Recibió el doctorado en ciencias en computación en 2020 por el CINVESTAV-IPN. Actualmente, es profesor investigador de tiempo completo en el ITESM, Campus Monterrey, y miembro del SNII-CONAHCYT nivel candidato. Su área de especialidad es la optimización evolutiva multiobjetivo en la cual cuenta con más de 25 publicaciones incluyendo artículos en revistas y congresos internacionales. Actualmente, dirige dos tesis doctorales, tres tesis de maestría y tres trabajos terminales a nivel licenciatura. Además, participa en un proyecto CONACyT "Ciencia de Frontera" sobre la búsqueda de arquitecturas neuronales óptimas para restauración de imágenes. Teléfono: +52 55-4188-6633. Correo electrónico: jfalcon@tec.mx.

Firma: