*Evaluación de la confiabilidad de las CNN implementadas a alto nivel.*

Moya Africano Jaime Andres  
Ingeniería Electrónica  
Universidad Pedagógica y tecnológica de Colombia Sogamoso   
[jaime.moya@uptc.edu.co](mailto:jaime.moya@uptc.edu.co)

PhD. Perez Holguín Wilson Javier   
Ingeniería Electrónica  
Universidad Pedagógica y tecnólogica de Colombia Sogamoso   
[wilson.perez@uptc.edu.co](mailto:Wilson.perez@uptc.edu.co)

Cristancho Sanchez Halder Mauricio  
Ingeniería ElectrónicaUniversidad Pedagógica y tecnólogica de ColombiaSogamoso  
[halder.cristancho@uptc.edu.co](mailto:halder.cristancho@uptc.edu.co)

PhD(C). Mesa Mesa Luis Ariel  
Ingeniería Electrónica  
Universidad Pedagógica y tecnólogica de Colombia Sogamoso   
[ariel.mesa@uptc.edu.co](mailto:ariel.mesa@uptc.edu.co)

***Resumen*—** Este paper review se centra en identificar y recopilar investigaciones pertinentes en el ámbito de la confiabilidad y robustez de las redes neuronales convolucionales. Gracias al empleo de gestores de búsqueda especializados, hemos logrado encontrar trabajos científicos que exploran aspectos cruciales, tales como la inyección de fallos en la inferencia y las estrategias de pruning. Asimismo, la aplicación de la herramienta VOSviewer nos ha brindado una perspectiva visual detallada de la red temática que enlaza a nuestros temas de interés, incluyendo las redes neuronales convolucionales, el proceso de pruning y los métodos de inyección de fallos en la inferencia a través de software.

***Palabras clave—*** *CNN, Pruning, SEU, modelo de falla, entrenamiento, inferencia, confiabilidad.*

# **Introducción.**

Las redes neuronales convolucionales (CNN) es una de las arquitecturas de Deep learning (aprendizaje profundo). “Inicialmente introducidas por Yann LeCunet et al. en 1998 a través de la red que llamaron Neocognitron”[1]. Se utilizan en una variedad de tareas especialmente para identificar patrones en imágenes, lo que las hace ideales para tareas como el reconocimiento de objetos y clasificación de imágenes.

Las CNN también son susceptibles a fallos, que pueden ser causados por una variedad de factores, incluyendo errores de programación, errores en los datos de entrenamiento, errores en la arquitectura de la red, errores de hardware y ataques maliciosos.

Los fallos pueden tener un impacto significativo en el rendimiento de la red. En algunos casos, los fallos pueden hacer que la red sea incapaz de realizar la tarea para la que fue diseñada.[2]

Esperamos que este proyecto permita conocer e identificar qué neuronas son más importantes al momento de la predicción de estas redes, así como también nos dará un panorama más claro de cómo mitigar estos fallos.

# **Metodología**

1. DATASET

El dataset en una CNN es fundamental para permitir que la red neuronal aprenda y mejore su capacidad de reconocimiento y extracción de características en imágenes. Este conjunto de datos no solo es esencial para el entrenamiento inicial de la red, sino que también se utiliza para afinar y optimizar su rendimiento a lo largo del proceso de validación y evaluación

El dataset se descompone en una matriz de píxeles, donde cada píxel representa un valor numérico que indica la intensidad del color en una ubicación específica es importante resaltar que un dataset debe contener etiquetas de las imágenes con las categorías o clases correspondientes.

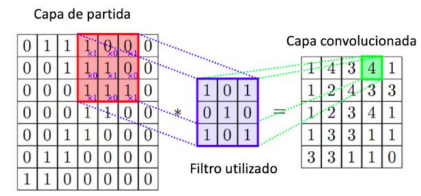
Un dataset debe ser suficientemente variado y representativo de las posibles entradas que la red podría enfrentar. Dentro de los dataset más utilizados se encuentra modified national institute of standards and technology (MNIST) y extended modified national institute of standard and technology (EMNIST)

1. ARQUITECTURAS

Las arquitecturas de las CNN están compuestas por múltiples capas que aplican filtros al conjunto de datos de entrada (dataset) con el objetivo de propagar los resultados hacia las capas subsiguientes. Esta técnica permite la detección y extracción de características relevantes en los datos, lo que contribuye al proceso de aprendizaje y a la representación efectiva de patrones en la red. Cada capa realiza operaciones de convolución y agrupación para capturar detalles importantes y reducir la dimensionalidad, permitiendo que la red aprenda de manera progresiva y abordar tareas complejas de manera eficiente.

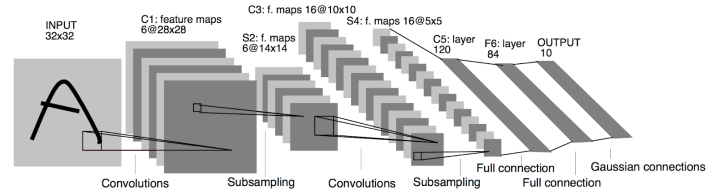
Como punto de partida la CNN toma como entrada son píxeles de una imagen, esta imagen puede ser en escala de grises o color, es importante resaltar que si es a color se necesitaran tres canales distintos los cuales se denominan red,green y blue.

Las capas convolucionales de una red neuronal reciben una matriz de entrada y la dividen en subconjuntos más pequeños al operar matemáticamente con una matriz reducida llamada kernel o filtro. Este filtro se desplaza a lo largo de las neuronas de entrada de izquierda a derecha, generando así una nueva matriz de salida conocida como la matriz de activación. Esta matriz de activación se convierte en una nueva capa de neuronas ocultas en la red(refenciaClasificación de imágenes usando redes neuronales convolucionales en Pytho). La representación gráfica de este proceso de convolución en una Convolutional Neural Network (CNN) se ilustra en la figura 1.



*Figura 1: Proceso de convolución de una CNN.*

Una vez aplicada la convolución, se procede a realizar una operación de reducción, conocida como "pooling", con el propósito de disminuir la cantidad de neuronas antes de llevar a cabo una nueva convolución. Posteriormente, los resultados obtenidos se propagan a través de una serie de capas interconectadas entre sí. Esta secuencia culmina con la capa de clasificación, donde se determina la categorización en función de los valores resultantes de los cálculos anteriores.(referencia Aceleración de la fase de inferencia en Redes Neuronales Profundas con dispositivos de bajo coste y consum). En la figura 2 se observa la arquitectura de una CNN LeNet propuesto por LeCun.



*Figura 2: Estructura red neuronal convolucional LeNet.*

1. MODELO DE ENTRENAMIENTO

El entrenamiento de una red neuronal requiere dos aspectos fundamentales: en primer lugar, un conjunto de datos lo suficientemente amplio y diverso que abarque todas las características relevantes del análisis que se está realizando; en segundo lugar, se necesita una potente capacidad computacional para procesar y aprender de estos datos de manera efectiva.[3]

[1] A. Moreno, “Clasificación de imágenes usando redes neuronales en Python,” pp. 1–81, 2019.

Un modelo de entrenamiento es un algoritmo que se utiliza para entrenar un modelo de aprendizaje automático mediante un conjunto de datos, este aprende a asociar los datos de entrada con los datos de salida mediante un proceso de optimización.

1. REPRESENTACIÓN DE DATOS

Los datos en una CNN se representan como tensores multidimensionales que fluyen a través de capas convolucionales, capas de agrupación y capas totalmente conectadas, donde se extraen características y se realizan operaciones para lograr el objetivo de la red, ya sea clasificación, detección de objetos, entre otras tareas.

En las redes neuronales convolucionales, los datos se representan mediante tensores multidimensionales, también conocidos como matrices o arreglos. Un tensor es una estructura de datos que puede tener una o varias dimensiones.

Estos tensores pueden tener cualquier número de dimensiones, existen distintos tipos de tensores como lo son:

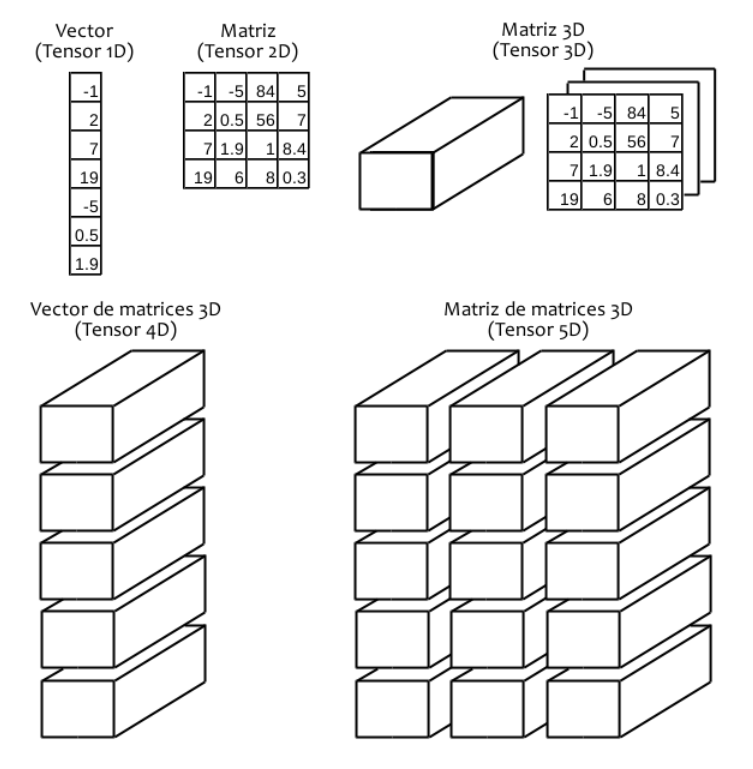
Los tensores de 1D se llaman vectores. Los vectores son una forma de representar datos que consiste en una lista de números.

Los tensores de 2D se llaman matrices. Las matrices son una forma de representar datos que consiste en una cuadrícula de números filas y columnas.

Los tensores de 3D se llaman cubos. Los cubos son una forma de representar datos que consiste en una colección de matrices.

En las redes neuronales, los tensores se utilizan para representar los datos de entrada, los datos de salida y los pesos de las neuronas. Los datos de entrada suelen ser representados por un tensor de 2D, los datos de salida suelen ser representados por un tensor de 1D, y los pesos de las neuronas suelen ser representados por un tensor de 3D.

En la figura 3 se pueden observar las diferentes representaciones gráficas de los tensores.



*Figura 3: Representación gráfica de los tensores.*

A su vez estos datos pueden tener una representación numérica tipo Floating-Point, Fixed-Point (punto flotante o punto fijo) o incluso representación binaria.

El tipo de representación de datos que se utiliza para los pesos depende de la tarea que se quiera realizar. Por ejemplo, si la tarea requiere una gran precisión, se utilizarán números de punto flotantes. Si la tarea requiere una gran eficiencia, se utilizarán números enteros. Si la tarea requiere una gran velocidad, se utilizarán números binarios.

1. PRUNING (poda)

El pruning es una técnica que se utiliza para reducir el tamaño y la complejidad de las redes neuronales convolucionales (CNN). Consiste en eliminar los pesos de la red que tienen un impacto menor en la precisión de la predicción.

Existen dos enfoques principales para llevar a cabo el pruning en una red neuronal. El primero de ellos involucra el entrenamiento exhaustivo de la red hasta que se alcance la precisión deseada, momento en el cual se aplica una única operación de poda. En el segundo enfoque, denominado poda iterativa, se inicia con el entrenamiento de la red y se le sigue una fase de poda inicial de pequeña magnitud. A partir de aquí, se repite el proceso de entrenamiento y poda gradualmente, aumentando la cantidad de poda en cada iteración. Este ciclo se repite de manera iterativa hasta que se logre la precisión deseada en la red neuronal.

Este método de poda iterativa tiene la ventaja de permitir que la red neuronal continúe aprendiendo y adaptándose durante el proceso de poda, lo que podría resultar en un modelo final más optimizado. Además, puede contribuir a evitar la pérdida abrupta de precisión que podría ocurrir con una única operación de poda(referencia An Incremental Pruning Strategy for Fast Training of CNN Models)

1. EVALUACIÓN DE CONFIABILIDAD (GOLDEN MODEL VS INYECCIÓN DE FALLOS)

El principal objetivo de este proyecto consiste en evaluar la confiabilidad de las redes neuronales convolucionales CNN mediante la inyección de fallos a los pesos más importantes de estas, todo esto mediante el método SEU (Single Event Upset). Se pretende analizar cómo los SEU afectan el rendimiento de la red y su capacidad para realizar tareas de clasificación de imágenes.

# **REVISIÓN PAPERS**

Iniciamos nuestra búsqueda explorando tesauros relacionados con las Redes Neuronales Convolucionales (CNN), en este proceso descubrimos una serie de términos claves que resultaron relevantes: "Convolutional Neural Networks," "Deep Neural Networks," "CNN," y "Convolutional Neural Nets." Estos tesauros nos brindaron una sólida base de conceptos y vocabulario esencial para profundizar en el estudio de las CNN y su aplicación en diversos campos. (meter referencia IEEE)

Siguiendo este enfoque, procedimos a plantear una serie de preguntas fundamentales en relación con nuestro tema de estudio:

¿Qué tan confiables son las redes neuronales convolucionales?

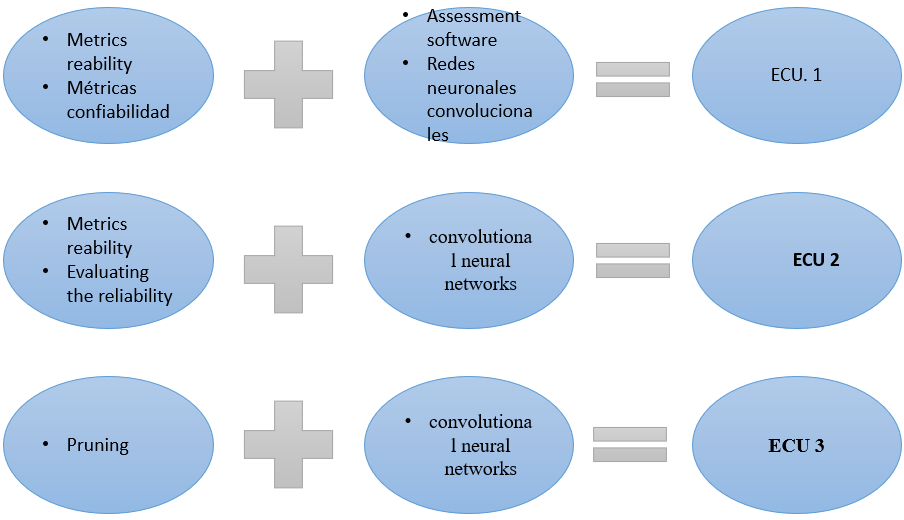
¿Cuál es el significado e importancia de las métricas de confiabilidad en las Redes Neuronales Convolucionales?

¿Qué métricas de confiabilidad se pueden aplicar en la inferencia de una red neuronal convolucional?

¿Cuáles son las métricas específicas de confiabilidad que pueden ser aplicadas al proceso de inferencia en las Redes Neuronales Convolucionales?

¿De qué manera se lleva a cabo la aplicación de inferencias en las Redes Neuronales Convolucionales?

Estas preguntas esenciales nos guiarán en el análisis profundo y en la exploración detallada de las métricas de confiabilidad y la aplicación de inferencias en el contexto de las Redes Neuronales Convolucionales, permitiéndonos crear ecuaciones de busque con el fin encontrar diferentes papers relacionados al tema de estudio. En la figura 4 se observan algunas de las ecuaciones de búsqueda utilizadas



*Figura 4: Ecuaciones de búsqueda.*

convolutional AND neural AND networks AND pruning

fault AND injection AND vulnerabilities AND in AND convolutional AND neural AND networks

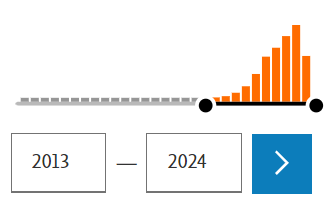
Una vez formuladas las ecuaciones de búsqueda, se procede a implementarlas en gestores de búsqueda especializados, como Google Scholar, PubMed, IEEE Xplore, arXiv y Elsevier Scopus, ScienceDirect. A través de estos recursos, se logra recopilar un considerable número de artículos científicos relacionados con el tema de estudio. Estas herramientas de búsqueda permiten explorar una amplia gama de fuentes académicas y bases de datos, extrayendo información relevante y actualizada que contribuirá de manera significativa a la investigación en curso

# **Análisis**

A través del proceso de investigación llevado a cabo, se ha constatado un marcado aumento en la cantidad de investigaciones relacionadas con redes neuronales convolucionales en los últimos años. Sin embargo, es relevante resaltar que gran parte de estas investigaciones se han centrado en abordar la inyección de fallos a nivel de hardware, dejando un espacio relativamente limitado para el análisis de la inyección de fallos en el ámbito del software. Este desequilibrio presenta un desafío considerable al buscar información que guíe hacia un análisis más profundo y completo en relación a la implementación de este tipo de fallas en software de alto nivel.

Cabe mencionar que la escasa información encontrada se centra principalmente en la inyección de fallos durante el proceso de entrenamiento de la red neuronal. En este sentido, se consideró esencial ampliar la búsqueda para explorar más a fondo la aplicación de fallos implementados en la inferencia. Esta ampliación de enfoque subraya la importancia de comprender los desafíos y riesgos potenciales asociados con la implementación de fallos en el software en el contexto específico de las redes neuronales convolucionales, lo que a su vez puede contribuir de manera significativa a la mejora de su robustez, confiabilidad y seguridad en aplicaciones prácticas.

La Figura 5 revela la notable cantidad de investigaciones realizadas sobre el tema de las redes neuronales convolucionales desde el año 2013 hasta la fecha actual. Es crucial resaltar que este estudio de interés ha experimentado un crecimiento exponencial a lo largo de los años, lo cual evidencia la creciente importancia y relevancia que las redes neuronales, en particular las CNN, están obteniendo en el ámbito científico. Este fenómeno refleja el reconocimiento de su potencial para abordar una variedad de desafíos y aplicaciones en campos que van desde el procesamiento de imágenes y reconocimiento visual hasta la medicina y la inteligencia artificial en general.

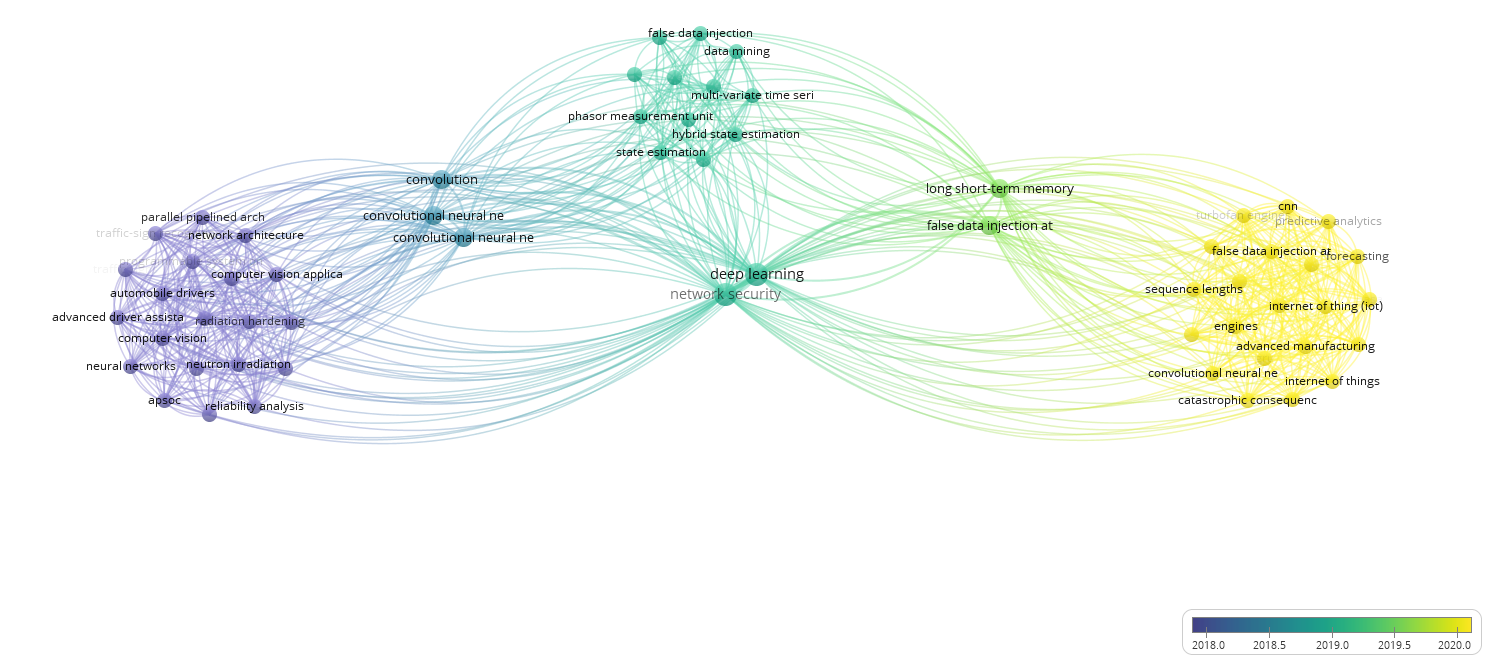
****

*Figura 5: Tendencia de papers acerca de CNN.*

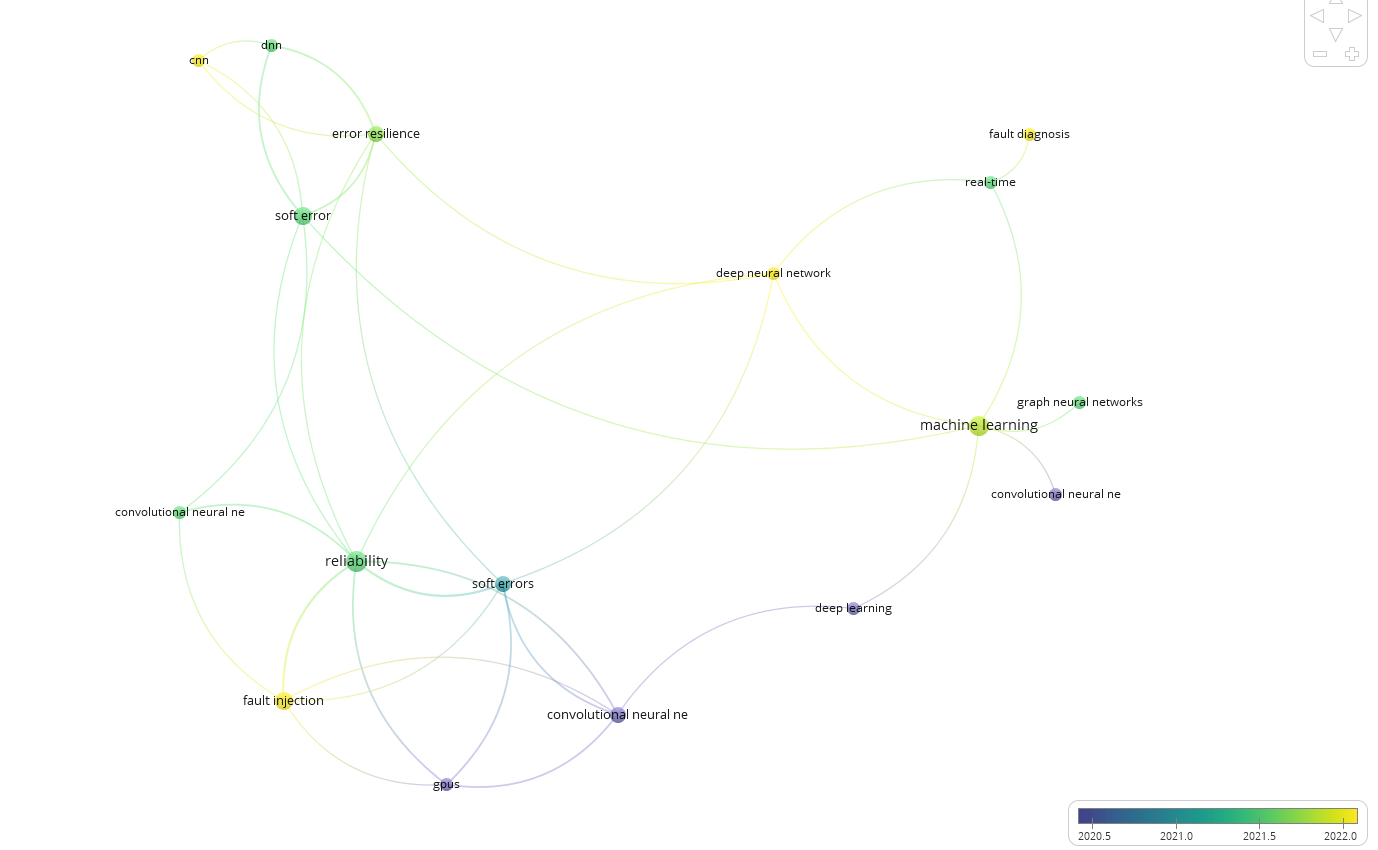
Al emplear la herramienta de análisis de visualización VOSviewer, se logra obtener un esbozo detallado y visualmente informativo de los temas interrelacionados que orbitan en torno al tema central de nuestro estudio. Esta herramienta permite explorar de manera sistemática y precisa la vasta red de conocimiento que se ha desarrollado en torno a las redes neuronales convolucionales y sus aplicaciones. Es posible identificar clústeres de investigación, conexiones relevantes entre conceptos, así como las tendencias emergentes y los autores influyentes en este campo.

El uso de VOSviewer se convierte en una herramienta esencial para trazar un panorama completo y comprensible de la evolución y la amplitud del conocimiento científico en torno a las redes neuronales convolucionales, lo que aporta una base sólida para nuestra propia investigación y análisis en este emocionante campo de estudio.

En la figura 6 se observa el amplio ámbito investigativo y científico que se presenta acerca de las CNN, mientras en la figura 7 se observa una profundización más cercana al tema relacionado con inyección de fallos y el pruning



*Figura 6: Tendencia de papers relacionados con CNN.*



*Figura 7: Tendencia de papers relacionados con inyección de fallos en CNN y uso de pruning.*

Uno de los métodos que se proponen es el uso del pruning esta puede ser una técnica eficaz para mejorar la confiabilidad de las redes neuronales convolucionales, permitiendo encontrar las neuronas más significativas en el aprendizaje de la red.[4]

El pruning es uno de los métodos previstos para la elección de las hotspots de la red neuronal, a pesar que se pueden diseñar métodos muchos más rústicos, este método por definición permite eliminar las neuronas menos influyentes en el entrenamiento de la red, ayudando a reducir la búsqueda de estos hot spots.

Sin embargo, es importante señalar que el pruning no es una solución perfecta. El pruning puede reducir el impacto de los fallos, pero no puede eliminarlos por completo. Además, este puede reducir la precisión de la red neuronal, especialmente si se realiza de forma desmedida.

El pruning puede utilizarse para mejorar la seguridad de las redes neuronales que se utilizan en aplicaciones críticas, como el reconocimiento facial, el reconocimiento de voz y el procesamiento del lenguaje natural, entre otros.

##### **Referencias**

[1] J. M. Aguilar, G. Botella, J. Alberto, and A. Del Barrio García, “Aceleración de la fase de inferencia en Redes Neuronales Profundas con dispositivos de bajo coste y consumo Inference acceleration in Deep Neural Networks on low-cost and low-consumption devices,” 2020.

[2] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, “EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples,” 2018. [Online]. Available: www.aaai.org

[3] L. M. Luza *et al.*, “Emulating the Effects of Radiation-Induced Soft-Errors for the Reliability Assessment of Neural Networks,” *IEEE Trans Emerg Top Comput*, vol. 10, no. 4, pp. 1867–1882, Oct. 2022, doi: 10.1109/TETC.2021.3116999.

[4] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, “Pruning Convolutional Neural Networks for Resource Efficient Inference,” Nov. 2016, [Online]. Available: http://arxiv.org/abs/1611.06440

[3] A. Moreno, “Clasificación de imágenes usando redes neuronales en Python,” pp. 1–81, 2019.