COLOMBIA POTENCIA DE LA VIDA

TIC

Adversarial Training

Entrenamiento adversarial como método de regularización

TALENTO TECH

AZ | PROYECTOS EDUCATIVOS

UTP Universidad Tecnológica de Pereira

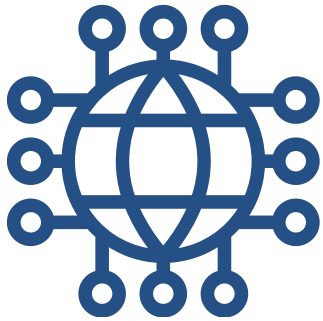
Adversarial Training: entrenamiento adversarial como método de regularización

El entrenamiento adversarial es un método de regularización en el aprendizaje profundo que se basa en la confrontación entre dos modelos: el generador y el discriminador. La idea central es establecer una competencia entre estos dos componentes para mejorar la capacidad del modelo y lograr una mejor generalización.

Conozcamos estos componentes:

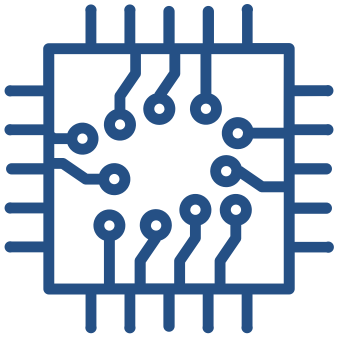
Modelo generador

Este componente crea datos sintéticos que se introducen en el conjunto de entrenamiento. Su objetivo es generar muestras que sean lo más parecidas posible a los datos reales.



Modelo discriminador

Este componente tiene la tarea de distinguir entre datos reales y datos generados por el generador. Su objetivo es volverse cada vez más preciso en esta tarea.



El proceso se realiza de la siguiente manera:

Conozcamos algunas de las situaciones:

1 **El generador crea datos sintéticos y los agrega al conjunto de entrenamiento.**	2 **El discriminador evalúa la autenticidad de todos los datos (reales y generados).**
3 **Ambos modelos se actualizan en función del rendimiento.**	4 **El generador ajusta su estrategia para generar datos más realistas, y el discriminador mejora su capacidad para diferenciar entre lo auténtico y lo generado.**

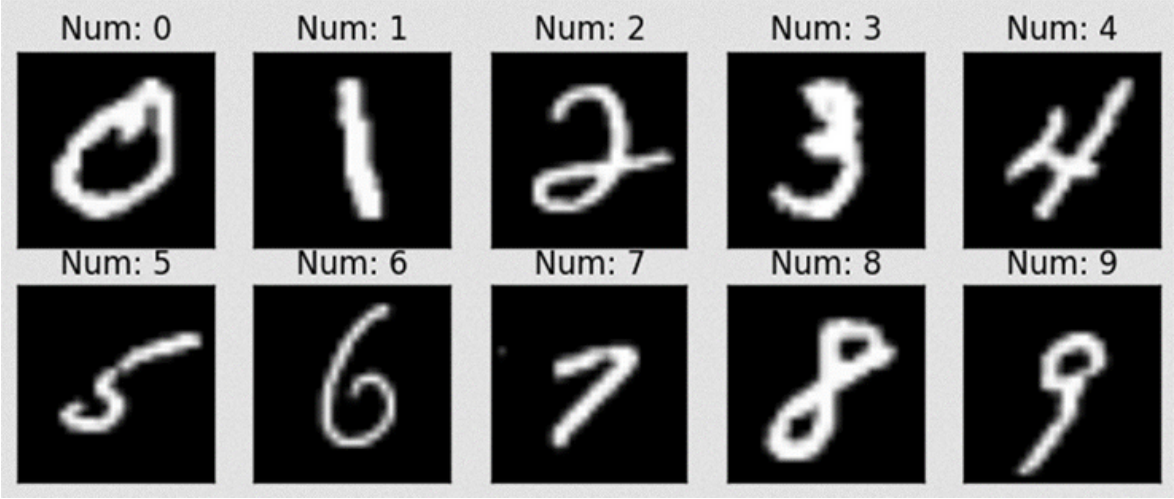
Este proceso de competencia iterativa lleva a una mejora continua en ambos modelos. El generador aprende a generar datos más realistas, y el discriminador se vuelve más experto en discernir entre datos auténticos y generados.

La principal ventaja de este enfoque es que ayuda a que el modelo generalice mejor a datos no vistos durante el entrenamiento, ya que se adapta a la verdadera distribución subyacente de los datos. Sin embargo, su implementación puede ser desafiante y requiere un ajuste cuidadoso de parámetros para evitar problemas como el sobreajuste.

Casos prácticos que ilustran cómo el entrenamiento adversarial mejora la robustez del modelo

• Caso 1: cómo el entrenamiento adversarial puede mejorar la robustez de un modelo de red neuronal en el contexto de reconocimiento de imágenes:

Suponga que está trabajando en un proyecto de reconocimiento de dígitos manuscritos utilizando una red neuronal convolucional (CNN). Su conjunto de datos incluye imágenes de dígitos escritos a mano, pero desea mejorar la capacidad del modelo para reconocer dígitos en condiciones adversas, como ruido o perturbaciones.



Implementación

Generador de Perturbaciones (Generador Adversarial):

1. Diseñe un generador adversarial que introduce perturbaciones controladas en las imágenes de los dígitos. Este generador toma una imagen de un dígito y agrega ruido o perturbaciones específicas para crear una versión perturbada.

Discriminador

2. implemente un discriminador que evalúa si una imagen perturbada es real (proveniente del conjunto de entrenamiento original) o generada por el generador adversarial.

Entrenamiento Iterativo

3. En cada iteración, alimente al modelo original (CNN) tanto con imágenes originales como con imágenes perturbadas generadas por el generador adversarial, el modelo original aprende a reconocer dígitos en condiciones adversas al ser expuesto a perturbaciones controladas.

Ajuste de Parámetros

4. Ajuste cuidadosamente los parámetros del generador adversarial y del modelo original para evitar el sobreajuste y garantizar una mejora en la robustez general.

Después de un entrenamiento iterativo, observará que el modelo original se vuelve más robusto frente a perturbaciones, como ruido o distorsiones en las imágenes. El generador adversarial ayuda a que el modelo aprenda características más robustas y generalice mejor a situaciones del mundo real.

Este ejemplo ilustra cómo el entrenamiento adversarial puede mejorar la capacidad de un modelo para manejar condiciones adversas y mejorar su robustez.