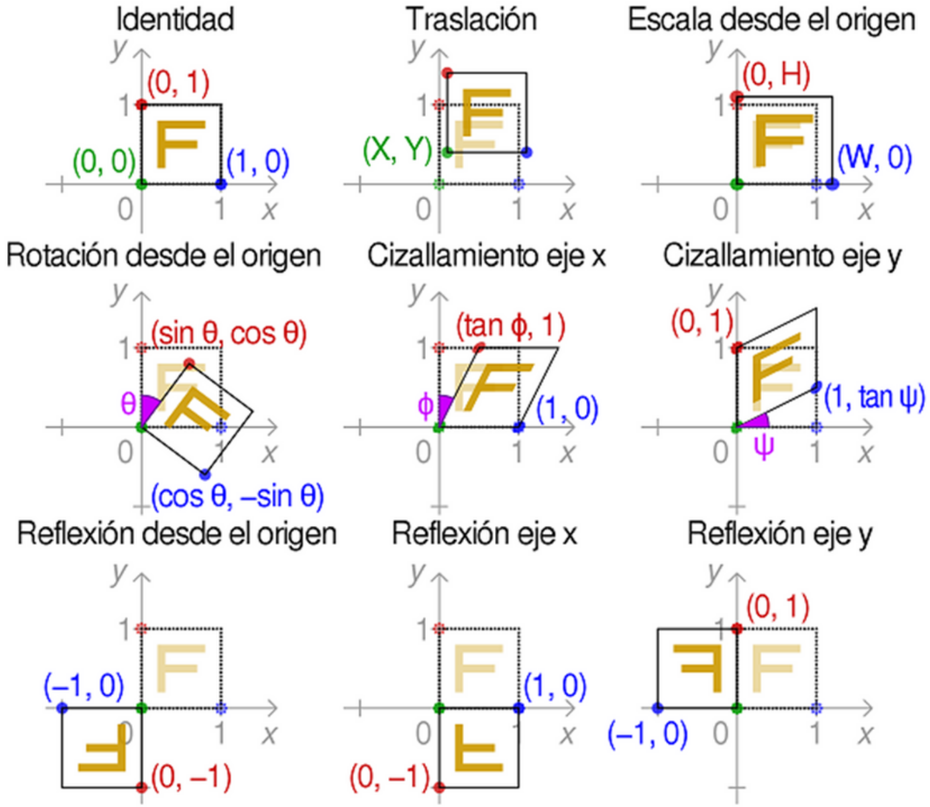
Aumento del conjunto de datos (Dataset Augmentation)

El Dataset Augmentation, es una técnica clave en el aprendizaje profundo que busca aumentar la diversidad y cantidad de datos disponibles para entrenar modelos. A menudo en la práctica, se puede presentar:

Los conjuntos de datos sean limitados en tamaño.	La diversidad no sea suficiente para que el modelo capture las variaciones del mundo real.

El Dataset Augmentation aborda este problema generando variaciones artificiales de las muestras existentes. A continuación, veremos algunas de las técnicas.



1. Rotación e inversión:

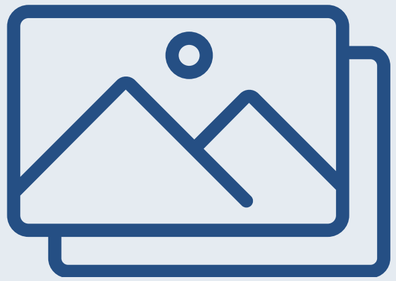
En aplicaciones de visión por computadora, rotar o invertir imágenes puede generar variaciones realistas. Por ejemplo, la imagen de un objeto en una orientación específica se puede rotar para simular diferentes ángulos de vista.

Implementación: utilizar funciones de bibliotecas como OpenCV o PIL para aplicar rotaciones e inversiones a las imágenes.

2. Desplazamiento y recorte:

Cambiar la posición de objetos dentro de una imagen o realizar recortes puede simular diferentes composiciones visuales.

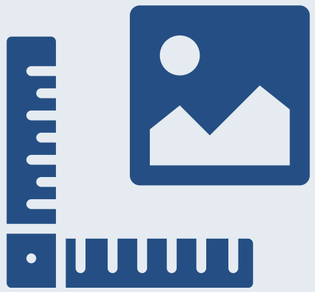
Implementación: ajustar la posición de las imágenes o recortarlas de manera aleatoria, manteniendo la información relevante.



3. Zoom y estiramiento:

Aplicar zoom o estiramiento a las imágenes puede simular variaciones en la distancia focal o en las proporciones.

Implementación: utilizar funciones que permitan cambiar el tamaño de las imágenes de manera controlada.



4. Cambio de color y brillo:

Variar la saturación, tonalidad y brillo puede hacer que el modelo sea más robusto a las variaciones de iluminación.

Implementación: ajustar los canales de color de las imágenes de manera aleatoria.

5. Transformaciones geométricas:

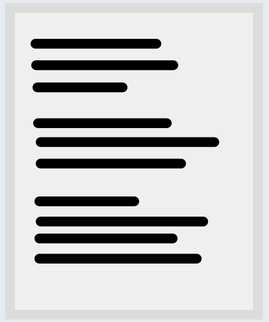
Aplicar transformaciones geométricas como distorsiones o perspectivas puede simular condiciones del mundo real.

Implementación: utilizar funciones que impliquen transformaciones geométricas a las imágenes.

6. Aumento en texto y secuencias:

En tareas de procesamiento de lenguaje natural, se pueden introducir variaciones en el texto como sinónimos, cambios gramaticales o truncamientos.

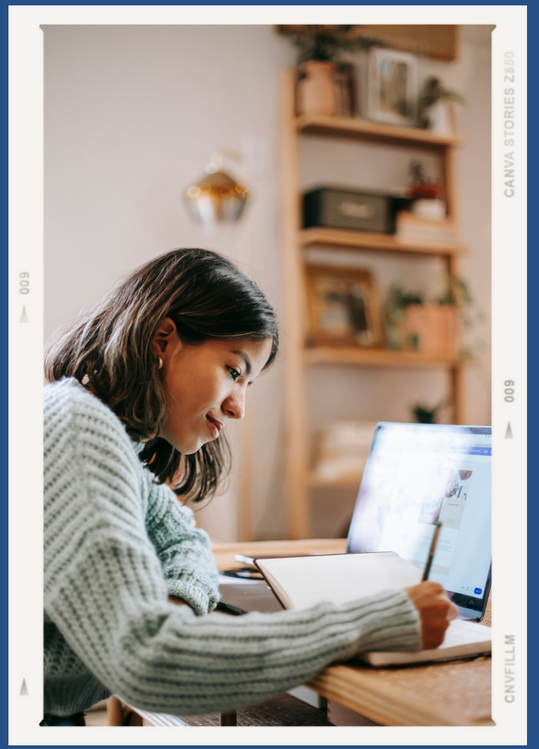
Implementación: utilizar técnicas de procesamiento de lenguaje natural para generar variaciones en el texto.



La clave del Dataset Augmentation es aplicar estas variaciones de manera coherente y mantener la integridad de las muestras originales; al incorporar estas técnicas, se enriquece el conjunto de datos de entrenamiento, mejorando la capacidad del modelo para generalizar a nuevas situaciones y mejorar el rendimiento en condiciones diversas del mundo real.

Ejercicio de demostración

Para demostrar las técnicas comunes de aumento de datos y su impacto en la generalización del modelo, se considera un escenario de clasificación de imágenes, utilizando un conjunto de datos limitado. Se le va a pedir a los estudiantes que creen un conjunto de datos de imágenes de gatos y perros, y se les pide entrenar un modelo para distinguir entre ambas categorías. Luego, aplicarán algunas técnicas de aumento de datos y observarán cómo afectan el rendimiento del modelo.



1

Conjunto de Datos Original:

Conjunto de datos limitado con imágenes de gatos y perros.

Cantidad de datos: 50 imágenes.

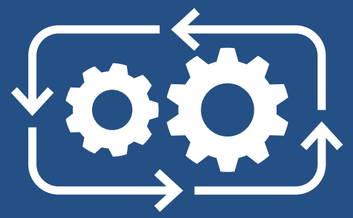
2

Aplicar a las imágenes las siguientes técnicas aumento de datos:

- Rotación e inversión
- Desplazamiento y recorte
- Zoom y estiramiento
- Cambio de color y brillo (saturación, tonalidad y brillo)
- Transformaciones geométricas y perspectivas

3

Procedimiento:



- Entrenarán dos modelos, uno sin aumento de datos y otro con las técnicas mencionadas.
- Medirán el rendimiento de ambos modelos en un conjunto de prueba independiente.

4

Impacto en la generalización:

- El modelo sin aumento de datos podría enfrentar problemas de sobreajuste debido al tamaño limitado del conjunto de datos original.
- El modelo con aumento de datos debería mostrar una mayor capacidad de generalización al haber sido entrenado con variaciones artificiales.

COLOMBIA POTENCIA DE LA VIDA

TIC

5

Evaluación del rendimiento:

- Compararemos las métricas de precisión, recall y F1-score en el conjunto de prueba para ambos modelos.

Este ejercicio práctico demostrará cómo las técnicas de aumento de datos pueden mejorar la generalización del modelo al proporcionar variaciones adicionales durante el entrenamiento.

TALENTO TECH

AZ | PROYECTOS EDUCATIVOS

UTP Universidad Tecnológica de Pereira