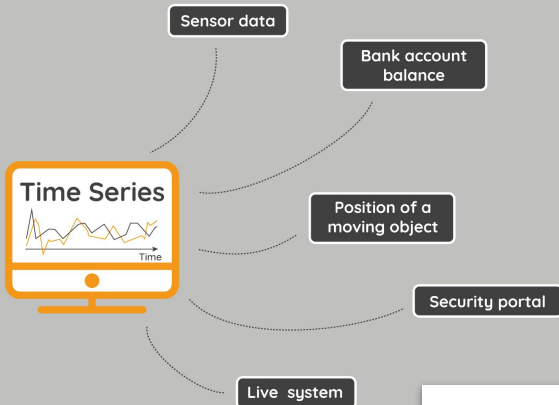


Preprocesamiento de Datos

Tipos de Datos

Tipos de Datos

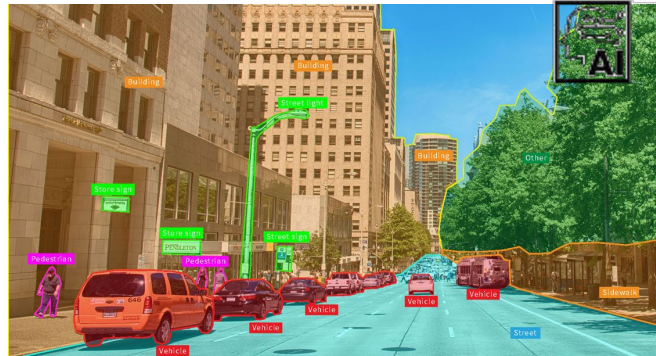
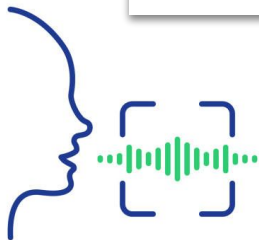
What is a time series?



Datos Estructurados

Series Temporales

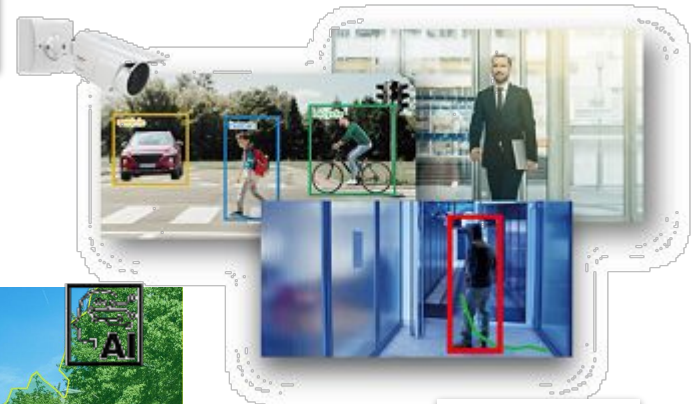
Audio/Voz



automated data mining survey responses computer transcripts qualitative data root cause classification insights ad-hoc analysis product reviews sentiment analysis of the customer dashboards consumer trends ad-hoc analysis early warning

text analysis

Texto



Vídeos

Imágenes

Preprocesamiento de Imágenes

Preprocesar Imágenes

1. Lectura de imágenes → Armado de datasets
2. Transformación de los datos
 - a. Escalado $[0, 255]$ a $[0, 1]$
 - b. Normalización
 - i. De $[0, 1]$ a $[-1, 1]$ → usando media y desviación igual a 0.5
 - ii. De $[0, 1]$ a $(-4, 4)$ → usando media y desviación del dataset
3. Carga de imágenes → **DataLoader**: envuelve un iterable en base al **Dataset** para permitir un fácil acceso a las muestras, mediante batch.
4. Data Augmentation

Escalado y Normalización de los datos

Escalado:

valor del pixel/255

$[0, 255] \rightarrow [0, 1]$

```
transform = transforms.Compose([transforms.ToTensor(),  
                                transforms.Normalize((0.1307), (0.3081))  
                                ])
```

Normalización:

$(\text{valor escalado del pixel} - \text{mean}) / \text{std}$

$[0, 1] \rightarrow [-1, 1]$

$[0, 1] \rightarrow [-4, 4]$

```
transform = transforms.Compose([transforms.ToTensor(),  
                                transforms.Normalize((0.5), (0.5))  
                                ])
```

```
transform = transforms.Compose([transforms.ToTensor(),  
                                transforms.Normalize((0.485, 0.456, 0.406), (0.229, 0.224, 0.225))  
                                ])
```

```
transform = transforms.Compose([transforms.ToTensor(),  
                                transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))  
                                ])
```

Normalización de Datos

¿Por qué normalizar los datos de imagen?

Hay varias razones por las que deberíamos normalizar los datos de imágenes:

- **Eliminar las diferencias de brillo y contraste entre diferentes imágenes**, que pueden deberse a variaciones en las condiciones de iluminación, la configuración de la cámara y otros factores.
- **Reducir el impacto de los valores atípicos**, que pueden sesgar la distribución de los datos y dificultar que el modelo aprenda de ellos.
- Ayudar a **acelerar el proceso de entrenamiento**, ya que el modelo puede converger más rápidamente cuando los datos de entrada son consistentes y predecibles.
- La normalización de los datos garantiza que el modelo reciba **entradas consistentes**, lo que facilita su entrenamiento y mejora su precisión.



Data Augmentation

ORIGINAL IMAGE



AUGMENTED IMAGE 1



AUGMENTED IMAGE 2



AUGMENTED IMAGE 3

Lectura de Datos

La [Clase ImageFolder](#) de PyTorch nos facilita crear conjuntos de datos de validación y entrenamiento. Pudiendo luego, usar este conjunto de datos para crear cargadores de datos iterables.

Para esto, debemos estructurar el directorio de los inputs de la manera como se muestra a la derecha.

El siguiente script nos permite realizar la lectura de las imágenes y carga del dataset particionado por train, validation y test; identificando cada input con la clase a la que pertenece.

```
# Cargar dataset de entrenamiento
train_dataset = datasets.ImageFolder(
    root='<path/to/folder/train>',
    transform=transform)
```

```
# Cargar dataset de entrenamiento
validation_dataset = datasets.ImageFolder(
    root='<path/to/folder/validation>',
    transform=transform)
```

```
├─ train
│   ├── class1
│   │   ├── 1.jpg
│   │   └── 2.jpg
│   └── class2
│       ├── 1.jpg
│       └── 2.jpg
├─ validation
│   ├── class1
│   │   ├── 1.jpg
│   │   └── 2.jpg
│   └── class2
│       ├── 1.jpg
│       └── 2.jpg
└─ test
    ├── class1
    │   ├── 1.jpg
    │   └── 2.jpg
    └── class2
        ├── 1.jpg
        └── 2.jpg
```

LET'S CODE

4. Pre-procesamiento de Datos: Dataset y DataLoader.ipynb



Preprocesamiento de Texto (semana 2)

PPT: [P. III - Procesamiento de Lenguaje Natural \(NLP\)](#)

Notebook: [6. Procesamiento de Lenguaje Natural](#)