



Introducción a la Ciencia de Datos

Maestría en Ciencias
de la Computación

Dr. Irvin Hussein López Nava



UNSURE IF BAD MODEL

OR INSUFFICIENT DATA

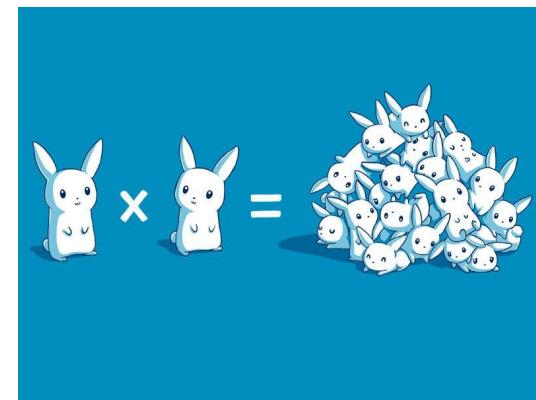
memegenerator.net

2.2 Aumento de datos

Let's keep talking (and listening)

about Data...

but augmenting



¿Qué es el aumento de Datos?



Definiciones

ChatGPT

Es una técnica en el aprendizaje automático que consiste en aplicar transformaciones a los datos de entrenamiento originales para generar nuevas muestras similares. Esto aumenta la diversidad de los datos y ayuda a los modelos a generalizar mejor y evitar el sobreajuste.

<https://chat.openai.com/>

Wikipedia

Es una técnica de aprendizaje automático que se utiliza para reducir el sobreajuste al entrenar un modelo de aprendizaje automático, entrenando modelos en varias copias ligeramente modificadas de datos existentes.

[https://en.wikipedia.org/wiki/
Data_augmentation](https://en.wikipedia.org/wiki/Data_augmentation)

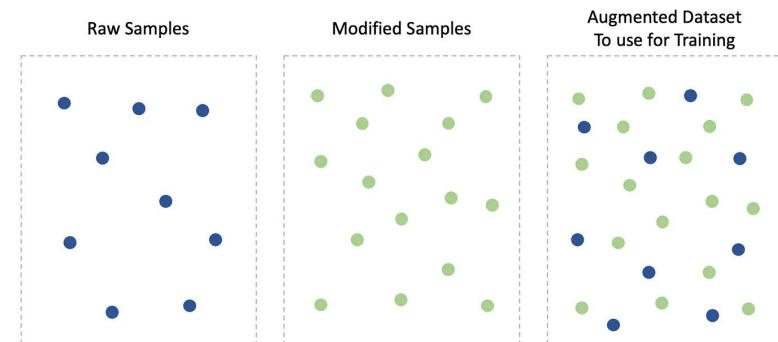
Gemini

Es una técnica utilizada en el aprendizaje automático que consiste en generar artificialmente nuevos datos a partir de un conjunto de datos existente. En otras palabras, es como hacer copias de tus datos pero con algunas modificaciones para que parezcan diferentes.

<https://gemini.google.com/>

¿Solo a partir de aprendizaje automático?

- El **aumento de datos** es una técnica para aumentar artificialmente el conjunto de **datos** (entrenamiento) mediante la creación de **copias modificadas** de un conjunto de datos **existente**.
- Incluye realizar pequeños cambios en el conjunto de datos o utilizar el aprendizaje para generar **nuevos** puntos de datos.



Datos aumentados vs Datos sintéticos

Los **datos aumentados** se obtienen a partir de los datos originales con algunos cambios menores.

En el caso del aumento de imágenes, se realizan transformaciones geométricas y del espacio de color (volteo, cambio de tamaño, recorte, brillo, contraste).

Los **datos sintéticos** se generan artificialmente sin utilizar el conjunto de datos original.

A menudo se utilizan DNNs (Redes Neuronales Profundas), GANs (Redes Generativas Adversariales), y Transformers.



**Mi jefe a
los 16**



**Yo de
23**



¿Cuándo debe usarse el aumento?

1. Para evitar que los modelos se “sobreajusten” a los datos originales.
2. El conjunto de entrenamiento inicial es demasiado pequeño.
3. Para mejorar la precisión del modelo.
4. Para reducir el coste operativo de etiquetado y limpieza del conjunto de datos sin procesar.

Nota 1. El **aumento de datos** puede realizarse en las distintas etapas del proceso de Ciencia de Datos.

Nota 2. No se limita a imágenes, también puede aplicarse a video, audio, otras series de tiempo, texto, etc.

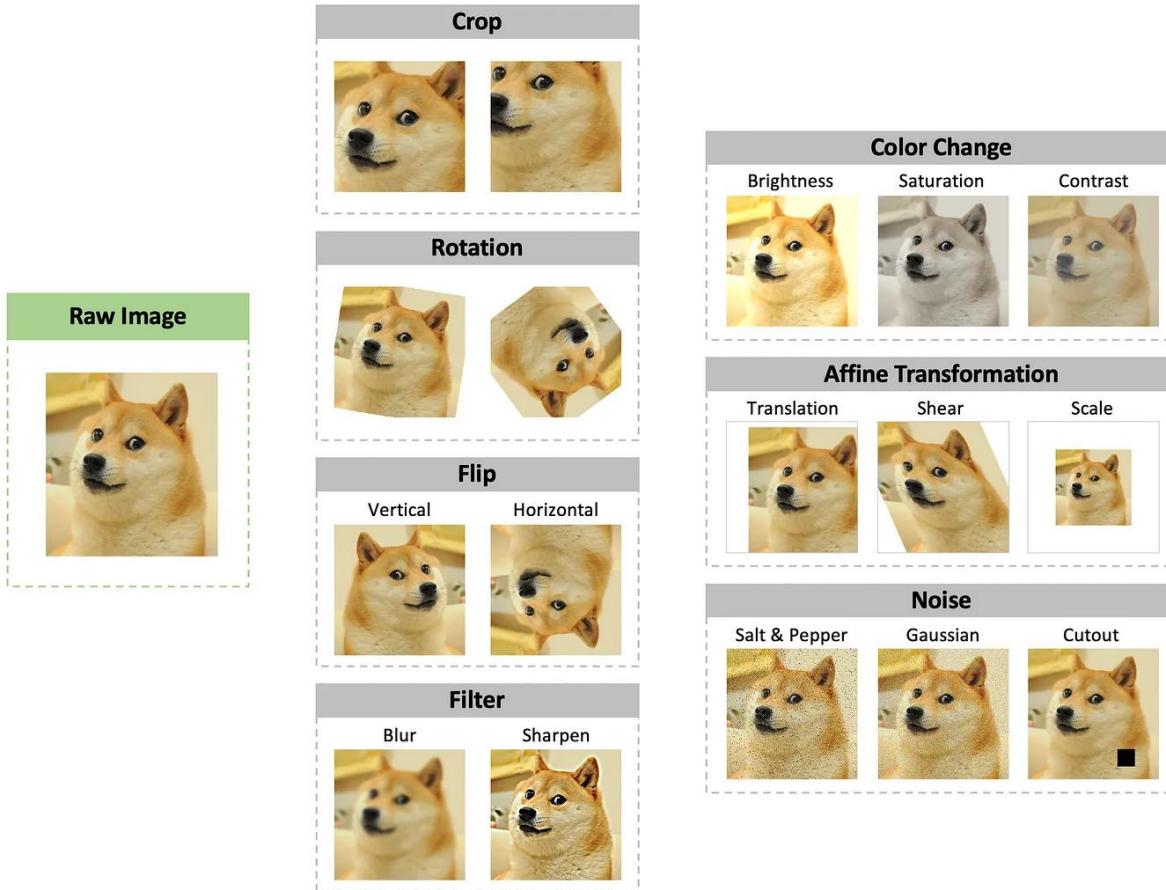
¿Cuáles serían las limitaciones?

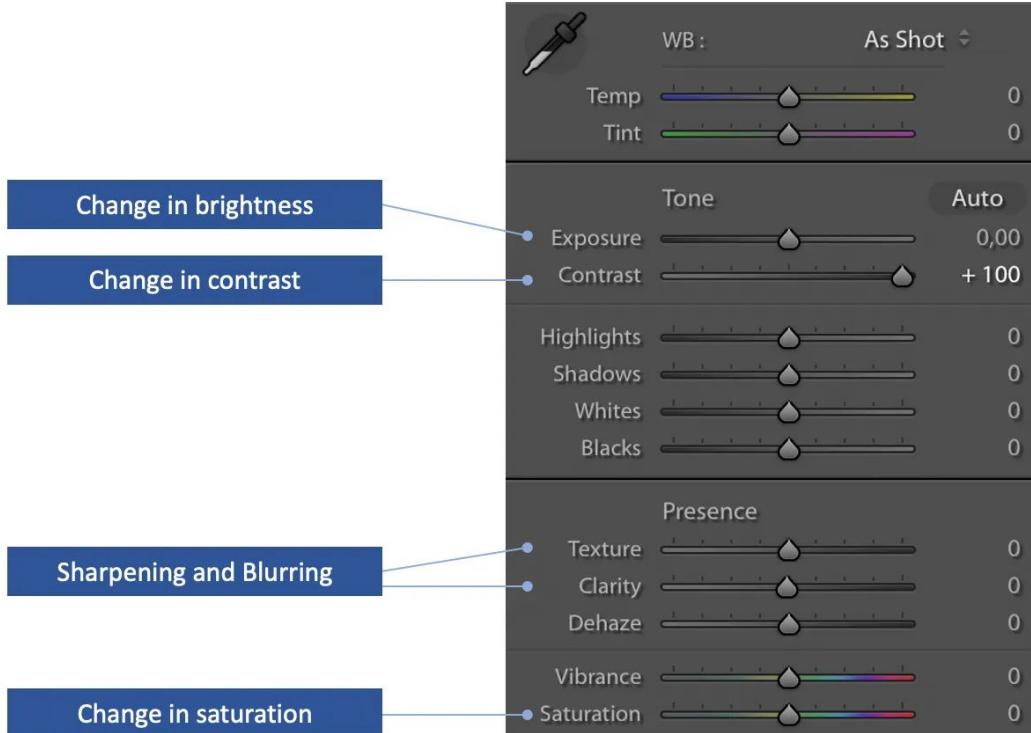
- Los **sesgos** del **conjunto de datos original** **persisten** en los **datos aumentados**.
- Garantizar la **calidad** del aumento de datos es costoso.
 - Igual de costoso si no se aplica adecuadamente.
- Se requiere investigación y desarrollo para construir técnicas avanzadas para cada tipo de datos.
 - Por ejemplo, la generación de imágenes de alta resolución mediante GANs puede resultar complicada.
- Encontrar un método eficaz de **aumento de datos** puede resultar complicado.

Aumento de imágenes

- **Transformaciones geométricas:** girar, recortar, rotar, trasladar, escalar, etc.
- **Transformaciones del espacio de color:** cambiar los canales de color RGB, el contraste, el brillo, entre otros.
- **Filtros de núcleo:** cambiar la nitidez o el desenfoque de la imagen.
- **Borrado e inserción:** borrar alguna parte de la imagen original, o insertar píxeles de forma aleatoria.
- **Mezcla de imágenes:** mezclar varias imágenes; originales o adicionales.

Nota: tener cuidado al aplicar múltiples transformaciones a las mismas imágenes, ya que esto puede reducir el rendimiento de los modelos.

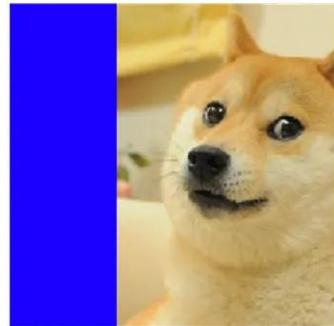




Translated Image



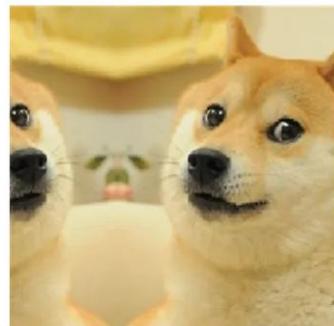
Fill Constant



Replicate



Reflect



Wrap

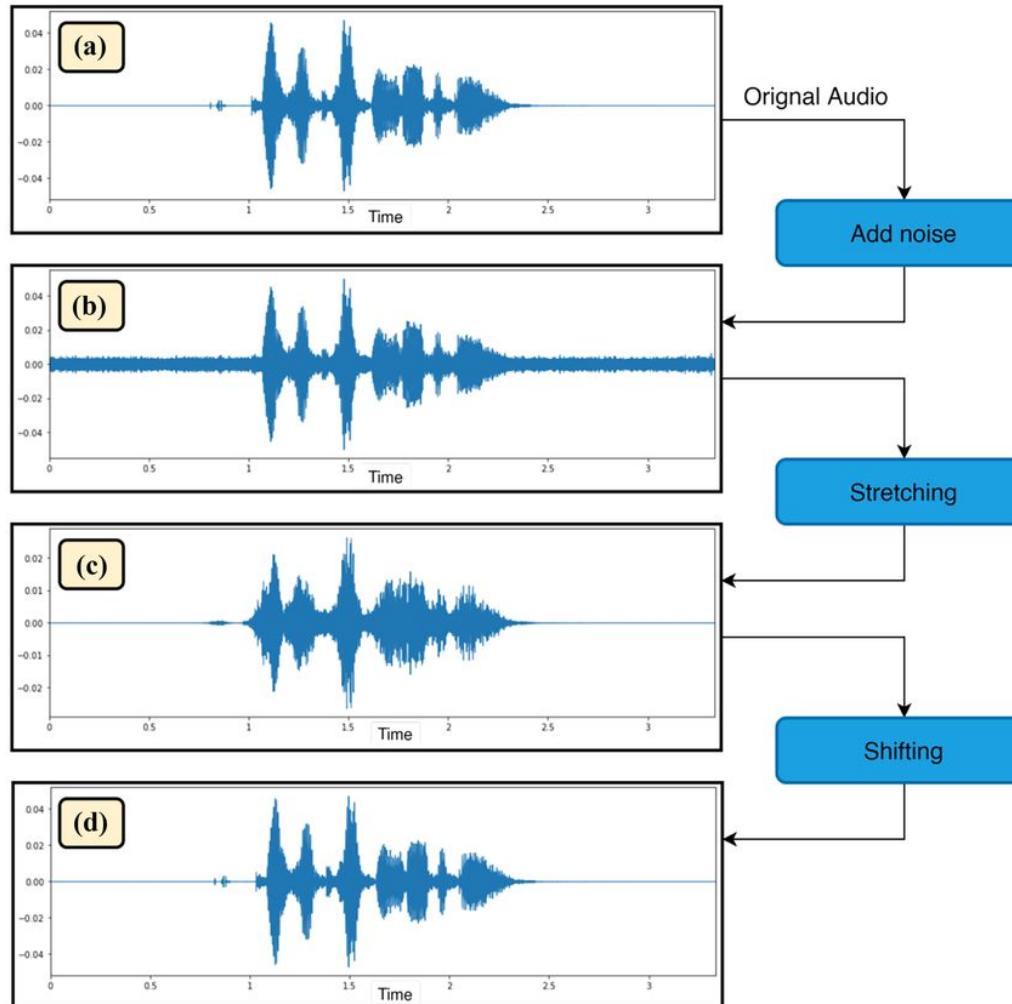


	Raw Image	Cropped Image	
Classification			Crops don't change labels in classification task. Part of the dog is still a dog.
			Crops do change labels in object detection task. Dog location (bounding box) after the crop changes.

Aumento de audio

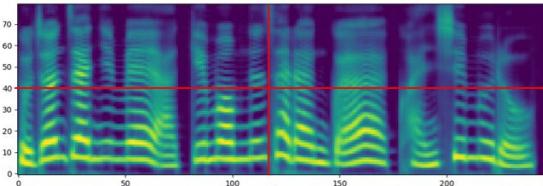
- ***Noise injection***: añade ruido gaussiano o aleatorio a las señales originales.
- ***Shifting***: desplaza el audio a la izquierda (*fast-forward*) o a la derecha con segundos aleatorios.
- ***Changing the speed***: 'estira' las series temporales a una velocidad fija.
- ***Changing the pitch***: cambia aleatoriamente el tono del audio.

Nota: los efectos de audio que se aplican digitalmente a la voz, o a los canales de música pueden considerarse aumento o síntesis de audio.

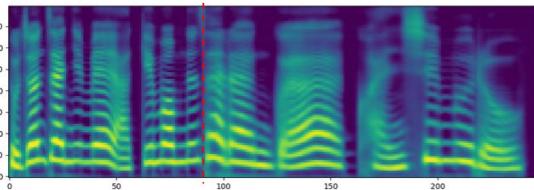




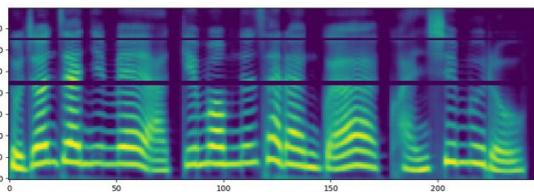
Original



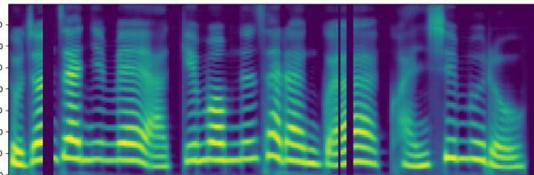
Time warping



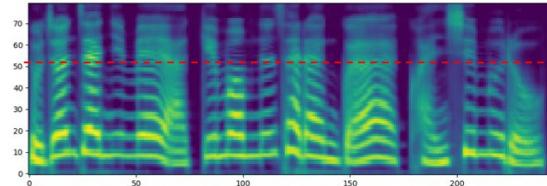
Frequency masking



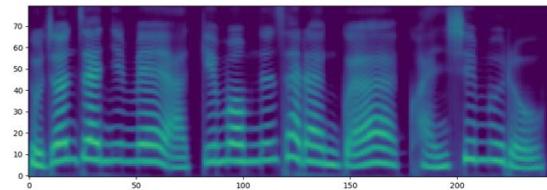
Time masking



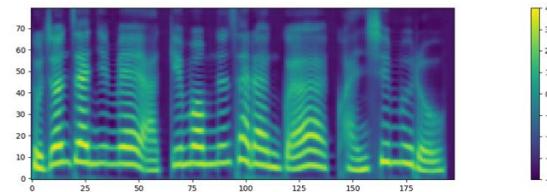
Frequency warping



Loudness control

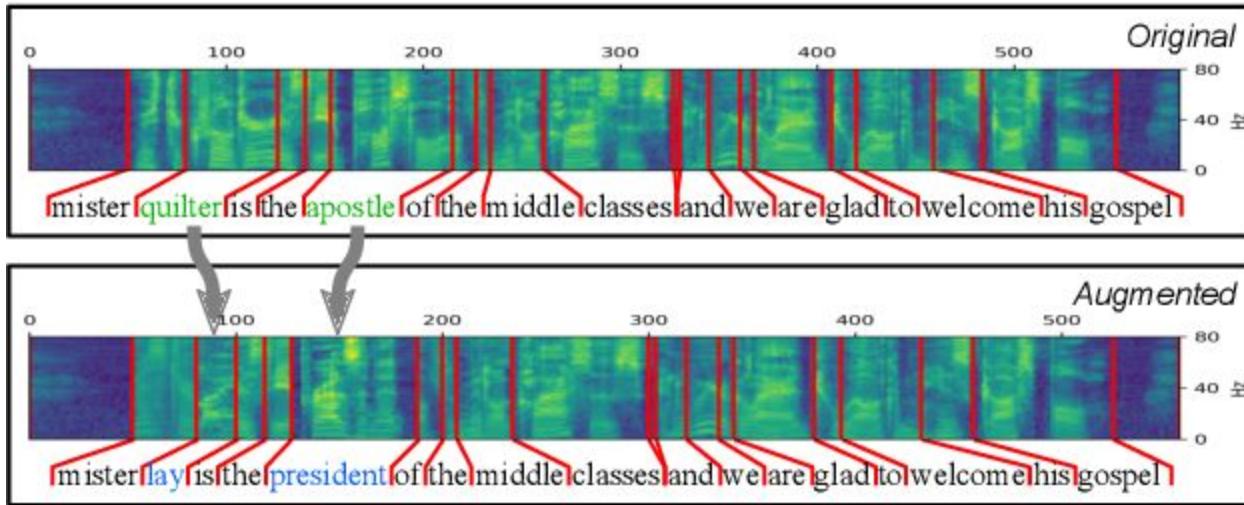


Time length control



Source line
Target line



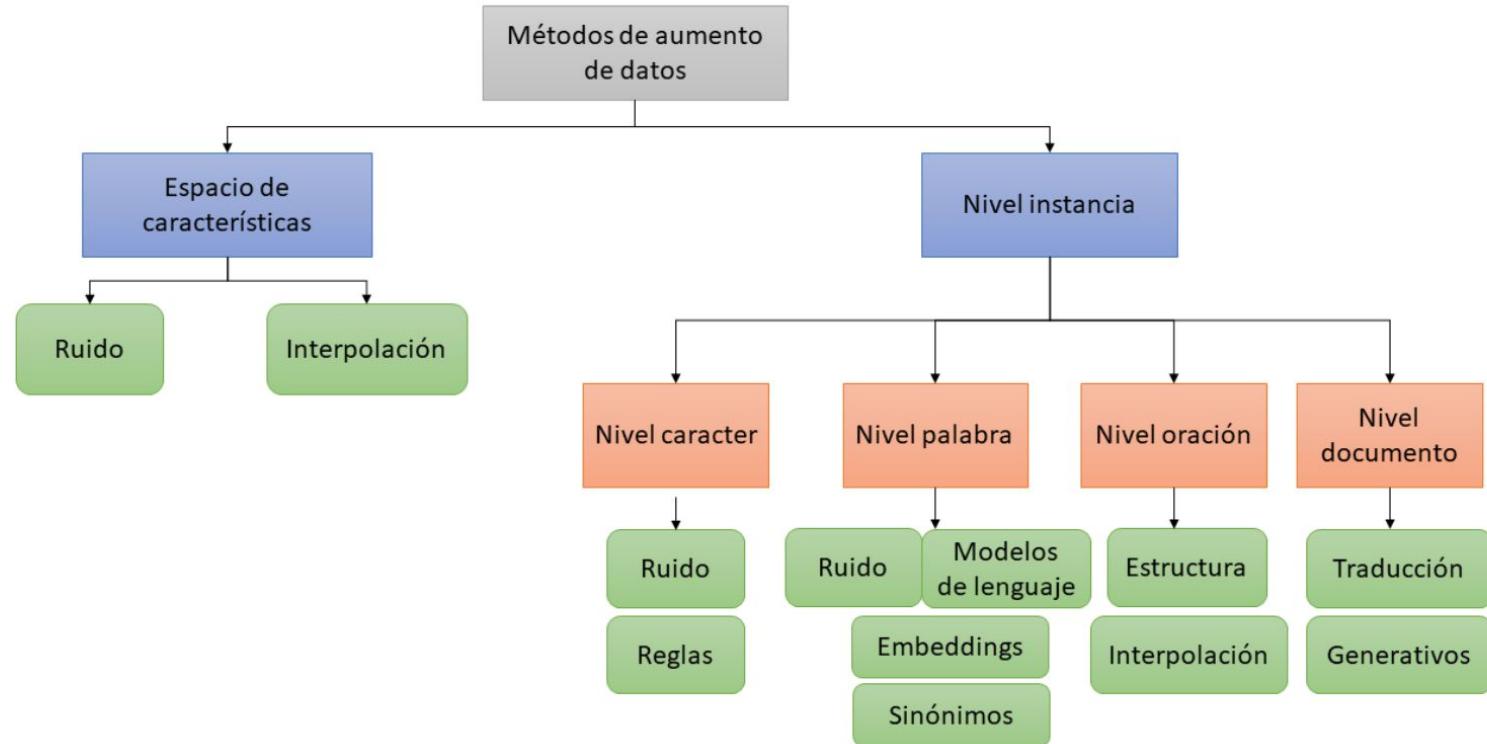


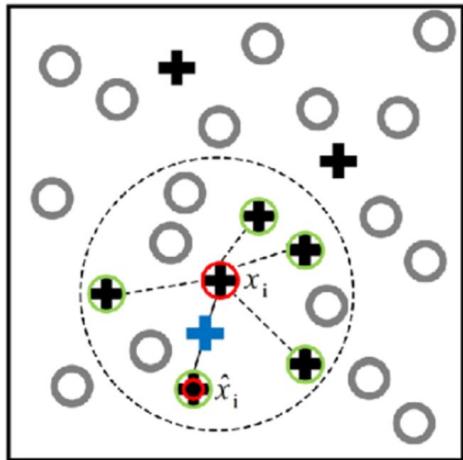
Aumento de texto

- **Word or sentence shuffling:** cambiar aleatoriamente la posición de una palabra o frase.
 - ¿por qué no cambiar letras?
- **Word replacement:** sustituir palabras por sinónimos.
 - Una técnica de esta familia es Back-translation
- **Syntax-tree manipulation:** parafrasear a nivel frase.
- **Random word insertion and deletion:** inserts words at random.

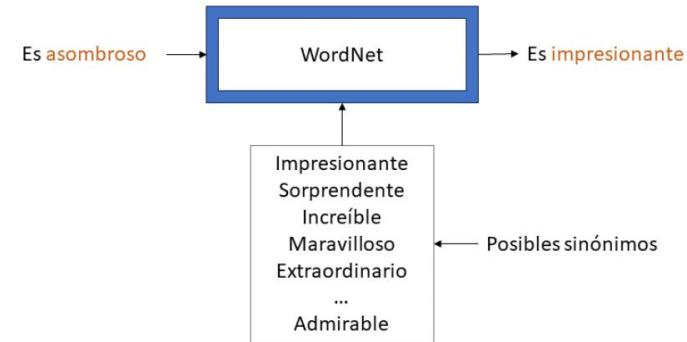
Nota: tomar en cuenta que los lenguajes naturales tienen reglas gramaticales, como la morfología, sintaxis, semántica y fonética.

-  Hasta mañana, compañeros.
-  Hasta mañana, compañeros.
-  Hasta mañana, compañeros.
-  Hasta mañana, compañeros.
-  Hasta mañana, compañeros.
-  Hasta mañana, compañeros.
-  ¡Vámonos que aquí espantan!





- Muestras de clase mayoritaria
- ✚ Muestras de clase minoritaria
- ✚ Muestra de clase minoritaria seleccionada al azar x_i
- ✚ 5 k-Vecinos más cercanos de x_i
- ✚ Muestra seleccionada al azar de los 5 vecinos \hat{x}_i
- + Instancia minoritaria sintética generada



Original:

Español

Anteriormente, el té se usaba principalmente para que los monjes budistas se **mantuvieran despiertos** durante la meditación.

Traducción

Francés

Auparavant, le thé était principalement utilisé par les moines bouddhistes pour rester éveillé pendant la méditation.

Nueva instancia:

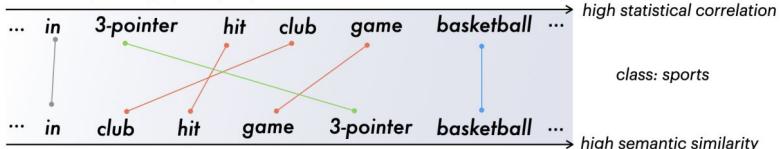
Español

En el pasado, el té se usaba principalmente para que los monjes budistas se **despertaran** durante la meditación.

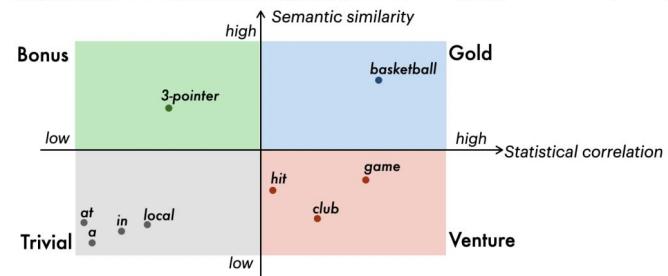
Traducción

"I hit a 3-pointer in a basketball game at a local club." (class: sports)

vocabulary ranking by two perspectives:



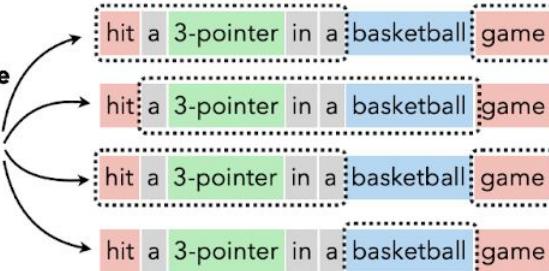
"I hit a 3-pointer in a basketball game at a local club." (class: sports)



Recognize word roles on original sample

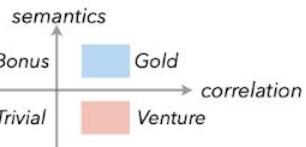
hit a 3-pointer in a basketball game

Select proper roles for augmentation

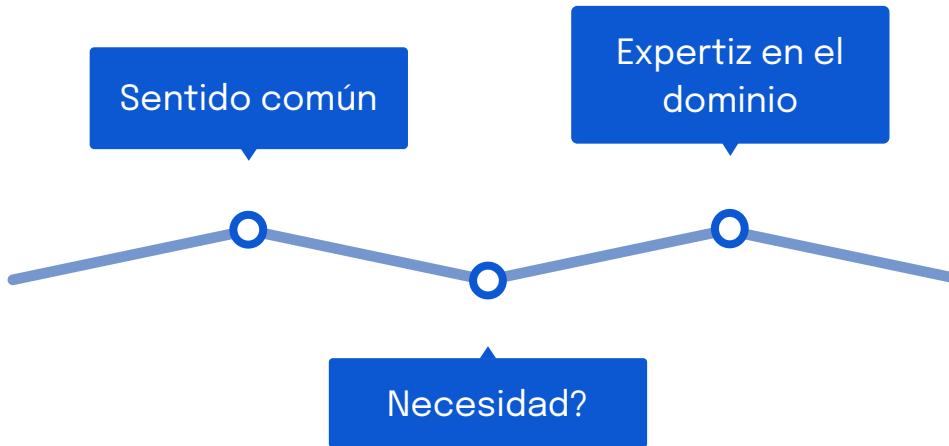


Augmented samples

candidate words



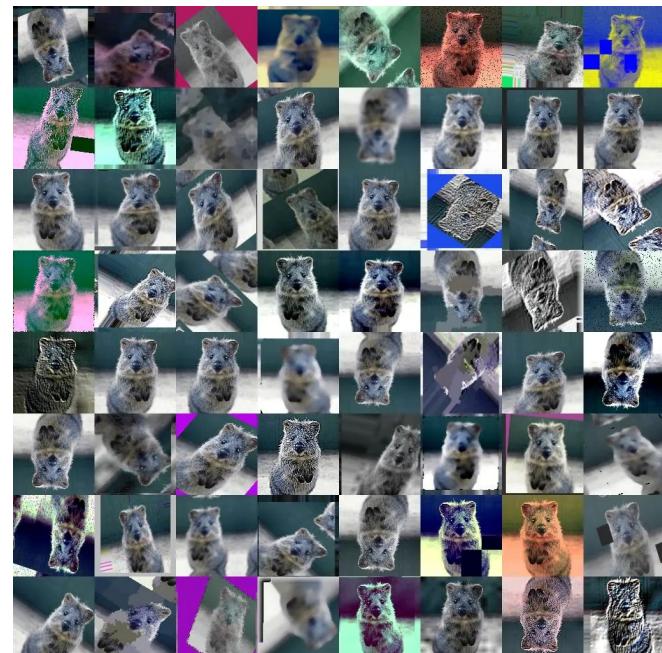
¿Cómo saber si aplicar o no el aumento?



Un vistazo en el dominio de las imágenes...

Primero, sentido común

- Existen demasiados tipos de **aumentos**.
- Hay que parar con las modificaciones antes de que la imagen se vuelva visualmente irreconocible.
- Si el ser humano no puede entender lo que hay en la imagen, ¿cómo podría hacerlo una máquina?





How much **data augmentation** is too much?

Necesidad

- Si trabajamos en conducción autónoma de autos. ¿Se deberían utilizar *Horizontal Flips* en para aumentar los **datos**?
 - Pues depende. ¿Se espera que el sistema vea imágenes al revés?

Raw Image

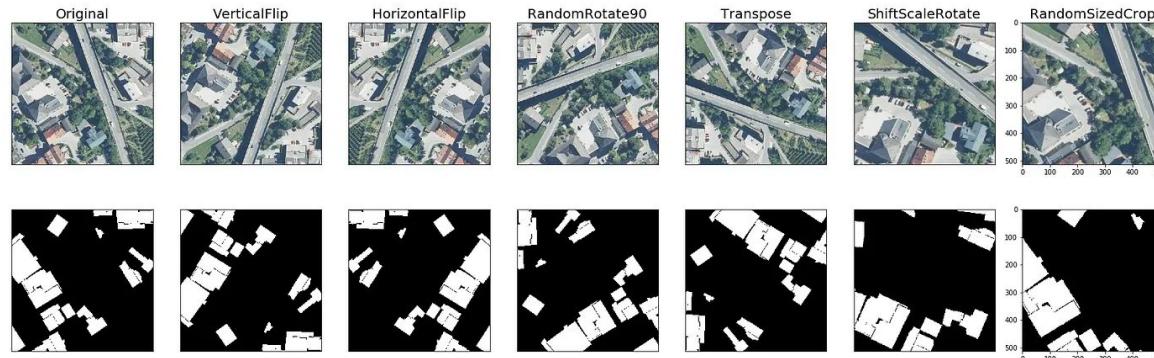


Horizontal Flip



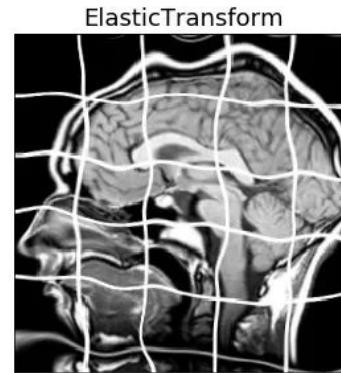
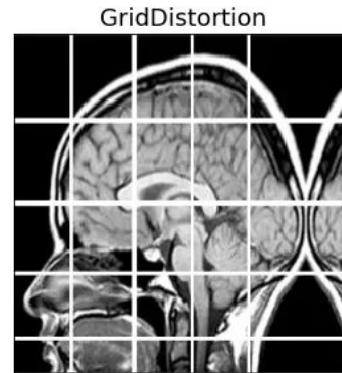
Expertiz en el dominio

- Dependiendo del ámbito de estudio, el **aumento de datos** tiene sentido o no.
- Por ejemplo, cuando se trabaja con imágenes satelitales, una buena elección para los aumentos sería el recorte, las rotaciones, los reflejos y la escala. Ya que no introducen distorsiones en objetos como edificios.



Expertiz en el dominio

- Por otro lado, cuando se trabaja con imágenes médicas, una mejor opción serían las transformaciones de color, ya que otras podrían no tener sentido.



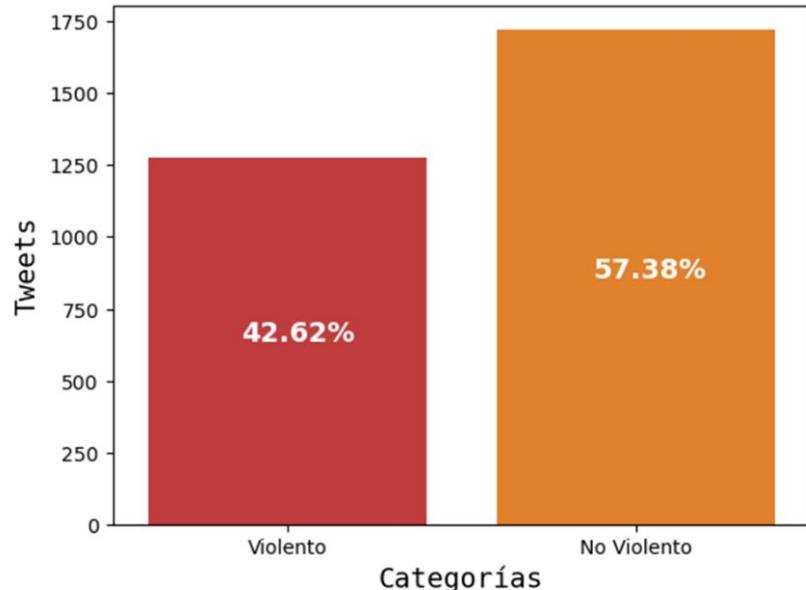
Caso de estudio

Eventos violentos

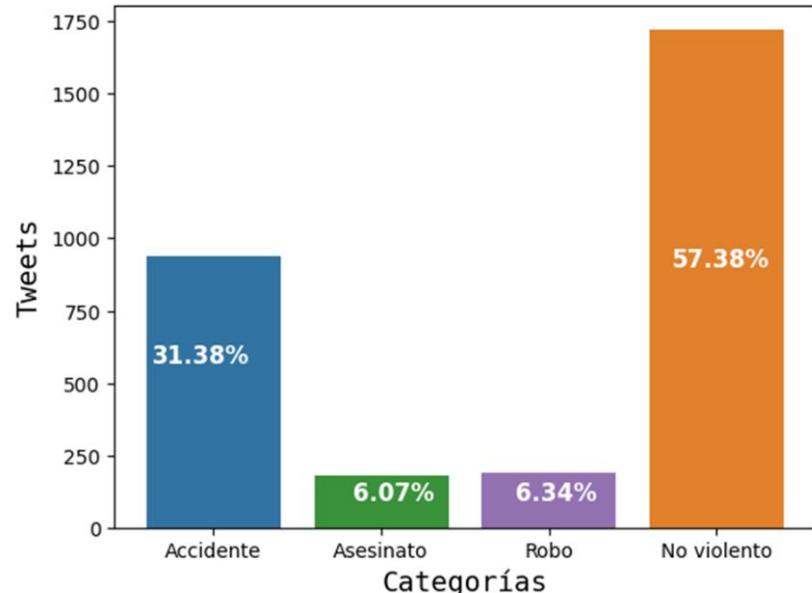
- La **violencia** tiene efectos negativos evidentes en quienes la presencian o experimentan, como una mayor incidencia de depresión, ansiedad y trastorno de estrés postraumático, entre otros.
 - Además, los sucesos **violentos** tienen un alto impacto para los gobiernos, ya que son los encargados de garantizar la seguridad a su población.
- Por ello, la detección y seguimiento de eventos relacionados con la violencia es fundamental.
 - En este contexto, las **redes sociales** constituyen una valiosa fuente de información para la detección y seguimiento de eventos **violentos**, ya que a menudo las personas publican la ocurrencia de eventos violentos en tiempo real.

Detección de eventos violentos en redes sociales

- El objetivo de la tesis ([Esteban Ponce](#)) fue diseñar un método que permitiera la identificación de publicaciones de eventos **violentos** en español y en Twitter,
 - A partir de información multimodal y técnicas de aumento de datos que mejoren el rendimiento de los métodos propuestos.
- El trabajo de investigación se dividió en dos fases experimentales: (i) identificación a partir de solo texto, (ii) a partir de texto e imágenes.
- La evaluación de los métodos se realizó utilizando los conjuntos de datos de DA-VINCIS 2022 y 2023.



(a) Proporción de las clases para la subtarea 1.



(b) Proporción de las clases para la subtarea 2.

Figura 28. Distribución de los datos para cada subtarea DA-VINCI 2023.

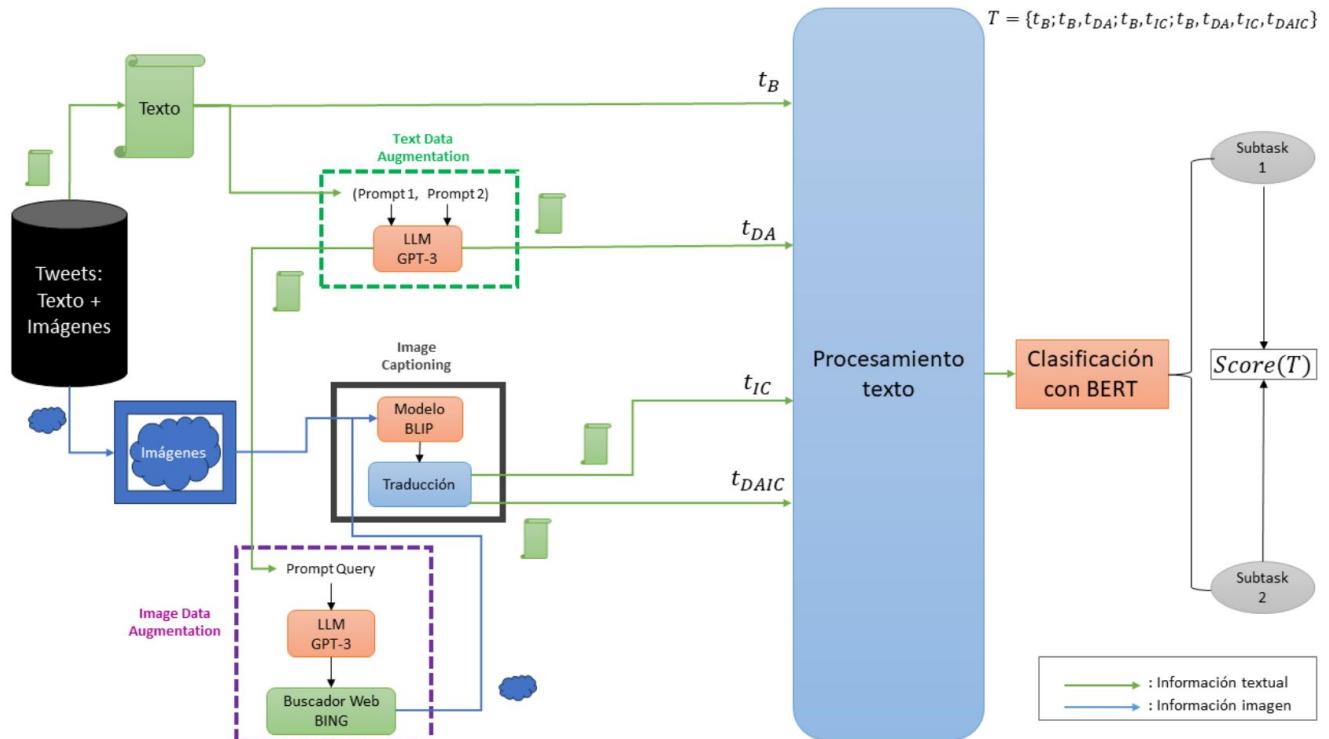


Figura 14. Proceso experimental seguido para resolver las subtareas propuestas en DA-VINCI 2023. Donde: T es la combinación de los textos; t_B es el texto original proveniente de los tweets; t_{DA} son los textos generados por GPT-3, t_{IC} son los textos obtenidos del proceso para obtener la descripción de imágenes y t_{DAIC} son los textos obtenidos del proceso de descripción de imágenes usando las imágenes recuperadas del paso de aumento de datos.

Prompt 1

"Write " + **number** + " different examples of tweets "+ **Tweet source** +" in spanish that reports different types of " + **crime + details** + " to different type of people "+ **place + country**"

```
Number = random(1,10)
country = " from the north region of Mexico"," from the south region of Mexico"," of Mexico", " of Latin America"," from the south of latin america"," from the north of latin america", "", "" "Spain"]
source = ["from the news", "from the authorities", "from the civilians", "", "from the victim", "from local news"]
place = "", " in different cities", " in different local stores", " in different streets", " in different avenues", " in different places"]
crime = "robberies", "homicide"
murder_details = [ "", " by guns ", " by robbery ", " by accident", " by fights", " by assaults ", "by assaults, accident, fights or detention of people for murder attempt", " or detention of people for murder"]
thief_details = [ "", " violent robberies", "assaults", "arrest for robbery", " attempts of robbery", " violent robberies, attempts of robbery, assaults ", " robberies, attempts of robbery, arrest of thieves"]
```

Ejemplo resultante:

Robo

La Fiscalía de Sinaloa informó sobre un ataque armado en la carretera Tepic-Guadalajara, donde dos turistas fueron asaltados por un grupo de criminales. #Asalto #Sinaloa

Asesinato

Una madre y su hija fueron asesinadas en Michoacán. #Homicidio #Michoacán #México

Figura 13. Prompt 1 utilizado para obtener nuevas instancias utilizando los modelos Davinci-003 de la familia GPT-3.

Prompt 2

Write **number** tweets in spanish about violent incidents related to **crime** that occurred in **country/region**. Add details such as time, location, what kind of robbery is (for example attempts, successful **crime**, with weapons, arrest, etc) and format the tweet according to what a **tweet source** would write. Feel free to add more or less details in order to make it more realistic. Limit the tweets to 265 characters at most.

```
Number = random(1,10)
country = " from the north region of Mexico", " from the south region of Mexico", " of Mexico", " of Latin America", " from the south of latin america", " from the north of latin america", "", "" "Spain"]
source = ["from the news", "from the authorities", "from the civilians", "", "from the victim", "from local news"]
crime = "robberies", "homicide"
```

Ejemplo resultante:

Thief

¡Atención! Robo a mano armada en la Calle Principal. Un individuo armado ingresó a una tienda y amenazó a los empleados y clientes. La policía se encuentra en busca del sospechoso. #RoboConArmas #Inseguridad

Murder

¡Terrible suceso! Se informa de un asesinato en la calle Insurgentes, Ciudad de México. Un ciudadano perdió la vida tras recibir múltiples disparos. Exigimos justicia y seguridad en nuestras calles. #JusticiaParaLasVíctimas #CDMX

Figura 16. Prompts utilizados para obtener nuevas instancias utilizando los modelos GPT-3.

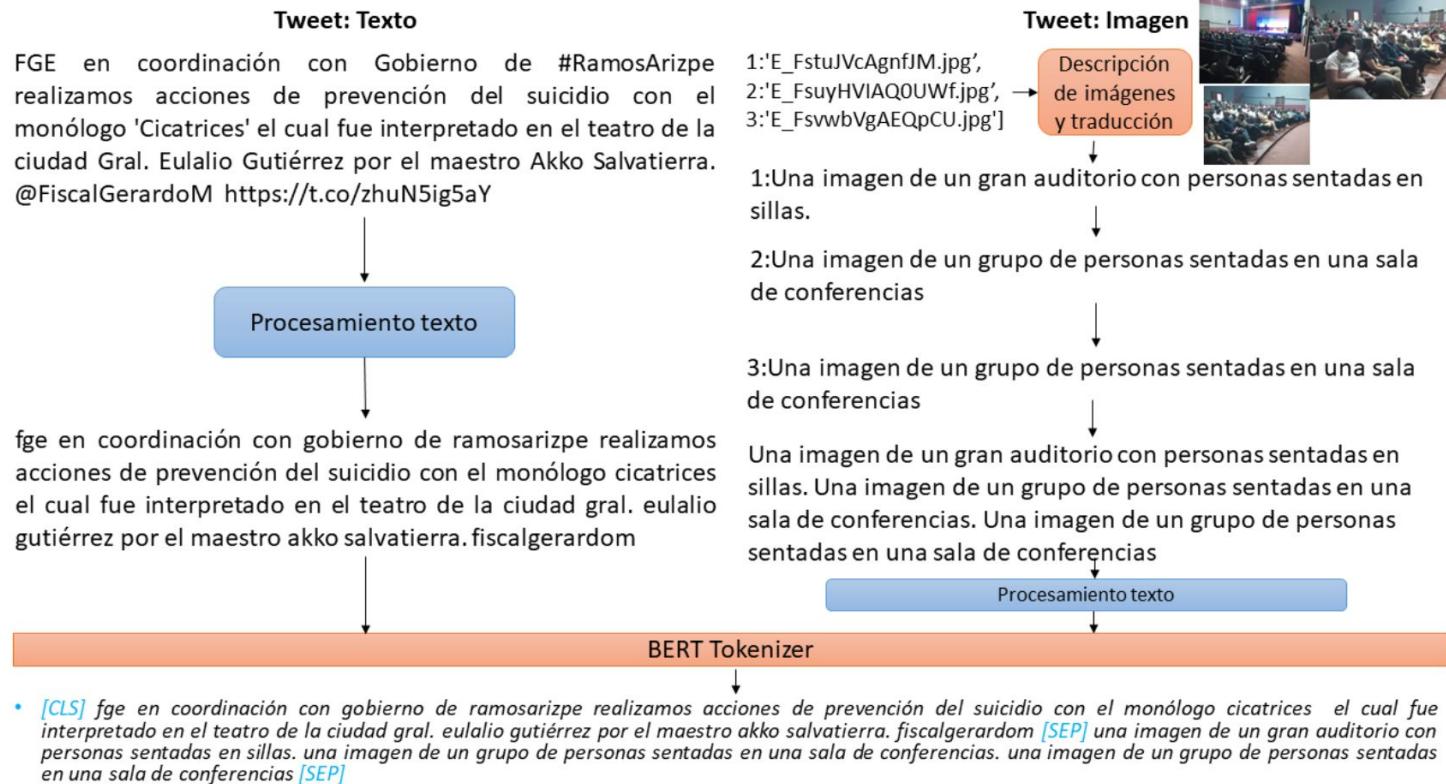
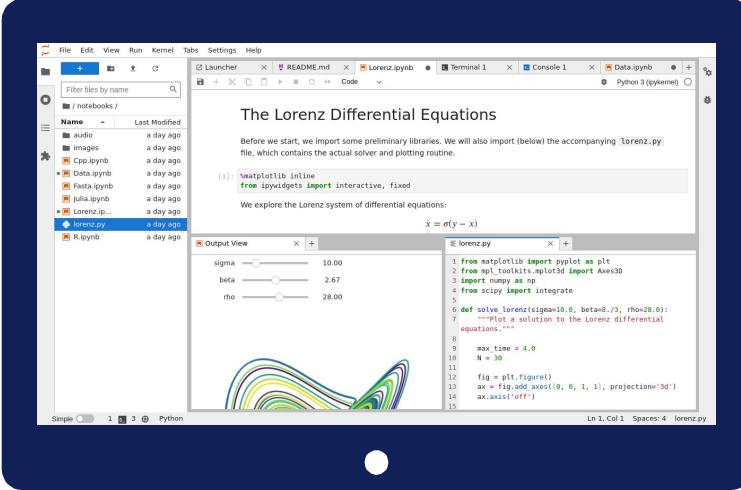


Figura 15. Ejemplo del resultado de aplicar el proceso de para obtener la descripción de imágenes.

Tweet: Texto	Query	Imagen
Una víctima fue dejada herida tras un asalto a una tienda en la Ciudad de México. #Asalto #Mexico	Asalto en la Ciudad de México	
La policía de Puebla reportó un ataque armado en el centro de la ciudad, donde dos hombres armados asaltaron a varios transeúntes. #Asalto #Puebla	Ataque armado Puebla	
¡Alerta! Robo a transeúnte en la Avenida Central. Una mujer fue víctima de un robo violento por parte de un ladrón en motocicleta. Mantengamos la precaución en espacios públicos.#RoboEnVíaPública #Cuidado	Robo a transeúnte Avenida Central	
¡Atroz crimen en la tienda de conveniencia de la colonia Roma! Una mujer de mediana edad perdió la vida en un intento de robo violento. Exigimos justicia y mayor vigilancia para garantizar la seguridad de los ciudadanos.#ViolenciaEnLasCalles#México	Crimen tienda conveniencia colonia Roma México	

Figura 17. Ejemplo de imágenes obtenidas al realizar una recuperación de imágenes basada en palabras clave.

(Go to live notebook)



The screenshot shows a Jupyter Notebook environment. On the left, there's a file browser with a list of notebooks and files. In the center, a code cell contains Python code for solving the Lorenz equations. Below the code cell is an output view showing three sliders for parameters sigma, beta, and rho, and a 3D plot of the Lorenz attractor. On the right, another code cell contains the source code for the `lorenz.py` file.

```
# Before we start, we import some preliminary libraries. We will also import (below) the accompanying lorenz.py file, which contains the actual solver and plotting routine.
%matplotlib inline
from ipywidgets import interactive, fixed

# We explore the Lorenz system of differential equations:
x = sigma*(y - x)
```

```
def lorenz(sigma=10.0, beta=10/3, rho=28.0):
    """Plot a solution to the Lorenz differential equations.

    Parameters
    ----------
    sigma : float
        The first parameter of the Lorenz system.
    beta : float
        The second parameter of the Lorenz system.
    rho : float
        The third parameter of the Lorenz system.
    max_time : float
        The time at which to stop the integration.
    N : int
        The number of points to plot.
    fig : matplotlib.pyplot.Figure
        A figure object to add the axes to.
    ax : matplotlib.pyplot.Axes3D
        An axes object to plot the solution on.
    """
    max_time = 4.0
    N = 30
    fig = plt.figure()
    ax = fig.add_axes([0, 0, 1, 1], projection='3d')
    ax.axis('off')
```

Gracias!

¿Alguna pregunta?

hussein@cicese.mx

<https://sites.google.com/view/husseinlopeznava>



CREDITS: This presentation was based on a template by [Slidesgo](#), and includes icons by [Flaticon](#).