Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Querétaro

Data Preparation

Autores:

A01368818 Joel Sánchez Olvera

A01661090 Juan Pablo Cabrera Quiroga

A01704076 Adrián Galván Díaz

A01708634 Carlos Eduardo Velasco Elenes

A01709522 Arturo Cristián Díaz López

TC3007C.501

Inteligencia artificial avanzada para la ciencia de datos II

Fecha:

30 de Octubre del 2024

Descripción de Datasets

Los datasets se encuentran en la siguiente carpeta de Drive.

Datasets de Bounding Box:

- Objetivo: Identificar exitosamente a las vacas en las imágenes, delimitando su contorno.
- Tamaño: 3950 imágenes.
- Tipo de Datos: Imágenes en vista aérea capturadas directamente sobre las camas de las vacas.

Dataset de Clasificación:

- Objetivo: Clasificar la posición de las vacas para determinar si están paradas o acostadas.
- Tamaño: 1472 imágenes.
- Tipo de Datos: Imágenes en vista aérea tomadas desde arriba, enfocadas en las vacas recortadas de las vacas.

Dataset de Clustering:

- Objetivo: Identificar patrones en la disposición y el uso de las camas de arena por las vacas.
- Tamaño: 9138 imágenes.
- Tipo de Datos: Imágenes aéreas centradas en las camas de arena.

Justificación de exclusión e inclusión de los datos

Para seleccionar las imágenes adecuadas en cada conjunto de datos, se establecieron criterios específicos en función de los objetivos de cada modelo. En el caso del modelo de bounding box, decidimos conservar las imágenes tal como fueron capturadas en condiciones reales, incluyendo aquellas con ruido, como imágenes oscuras o parcialmente obstruidas. Esta elección se realizó con el propósito de entrenar al modelo para que pueda manejar el ruido y las condiciones impredecibles de un entorno real, incrementando su robustez y capacidad de adaptación a situaciones diversas.

Para el modelo de clustering, se optó por mantener todas las imágenes disponibles sin ninguna exclusión. La razón detrás de esta decisión fue maximizar el volumen de datos

para identificar patrones en las camas de arena de forma más completa y precisa, favoreciendo la diversidad de muestras en diferentes condiciones.

Finalmente, para el modelo clasificador, fue necesario equilibrar las clases de las vacas en posición de pie y acostadas. Inicialmente, el modelo estaba generalizando demasiado bien las imágenes de vacas acostadas, afectando su precisión al clasificar vacas en posición de pie. Para abordar esto, se limitó la cantidad de imágenes de vacas acostadas en el conjunto de entrenamiento, igualando su proporción con las imágenes de vacas paradas. Este ajuste permitió mejorar la capacidad del modelo para diferenciar entre ambas posturas y ofrecer un rendimiento más equilibrado.

Limpieza de los datos

El único modelo que requirió un proceso de limpieza de datos fue el de clustering, donde se tomaron medidas específicas para optimizar el análisis de patrones en las camas de arena y asegurar que el modelo se centrara únicamente en el área relevante de cada imagen. En primer lugar, para evitar que el modelo identificara elementos irrelevantes fuera de la arena, se recortaron las imágenes para que solo mostraran el área de la arena. Este ajuste permitió que el modelo de clustering se enfocara exclusivamente en los patrones de las camas de arena, eliminando posibles distracciones en el resto de la imagen.

Adicionalmente, para evitar que el modelo generara clusters basados en las condiciones de iluminación en lugar de los patrones de la arena, el dataset se dividió en tres secciones según el brillo de la imagen, representando condiciones de día, tarde y noche. Esta estrategia permitió que las diferencias en la hora del día no influyeran en los clusters generados, ya que cada grupo de imágenes de brillo similar se analiza de manera independiente. Así, eliminamos la variable de iluminación como factor de clusterización y la utilizamos en cambio para separar el dataset en condiciones homogéneas de luz, favoreciendo la precisión del análisis.

Atributos y registros derivados

Para la preparación de los datos, empleamos la herramienta Roboflow para realizar el etiquetado de las imágenes, delimitando a las vacas presentes mediante bounding boxes. Este proceso permitió generar coordenadas precisas para cada vaca dentro de las imágenes, lo cual es esencial para entrenar el modelo de detección.

Cabe mencionar que hemos generado nuevos atributos basados en las coordenadas de los bounding boxes de las vacas en cada imagen, como las coordenadas (x,y) de las cajas. Hicimos esto ya que en el dataset original la ubicación y tamaño de cada vaca en las imágenes no estaban representados en los datos originales, por lo que fue esencial construir estos atributos derivados.

Dado que algunas imágenes contenían múltiples vacas, generamos registros derivados para representar a cada vaca individualmente en una imagen, creando así múltiples registros por imagen en muchos casos.