IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Fecha 17, Ene 2023.

Segmentación de Imágenes: Aplicados a un juego de tenis.

Omar Hernández[1], Luis Domínguez[1], Aldo Cervantes[1]
[1]Universidad Autónoma de Querétaro, Querétaro, Qro. México

RESUMEN Este artículo profundiza un ejemplo de segmetación de imágenes, explicando los dos tipos de segmentación, comparando y contrastando las diferencias entre la segmentación de instancia y semántica. Discutiremos los tipos de semánticas aplicados a un conjunto de datos que contiene imágenes de un juego de tenis etiquetadas por 2 clases, jugador de tenis y pelota. Obtenido de la plataforma Roboflow y aplicandolo en distintos modelos de segmentación

ÍNDICE DE TÉRMINOS Segmentación semántica, segmentación por instancias, Modelo fcn_8, Modelo vgg_16, Modelo vgg_unet

I. INTRODUCCIÓN

La segmentación de imágenes es la tarea de identificar y clasificar múltiples categorías de objetos.

Segmentación semántica

La segmentación semántica es una técnica que permite asociar a cada pixel de una imagen digital con una etiqueta de clase. También es considerada una tarea de clasificación de imágenes a nivel de píxel, ya que implica diferenciar los objetos de una imagen [1].

Es esencial entender que la segmentación semántica clasifica los píxeles de la imagen de una o más clases en ligar de los objetos del mundo real que no son semánticamente interpretables.

La segmentación semántica tiene como objetivo extraer las características antes de utilizarlas para formar categorías distintas en una imagen [2]. Las etapas son las siguientes:

- Analizar el conjunto de entrenamiento para clasificar un objeto especifico en la imagen.
- Crear una red de segmentación semántica para analizar los objetos y dibujar un bounding box alrededor de ellos.
- Entrenar la red de segmentación semántica para agrupar los pixeles en una imagen localizada creando una máscara de segmentación.

Algunas aplicaciones de segmentación semántica son en diagnósticos médicos, conducción autónoma, etc [3].

Segmentación por Instancias.

La segmentación por instancias es una forma única de segmentación de imágenes que ocupa detectar y delimitar cada instancia distinta de un objeto que aparece en una imagen. La segmentación por instancias detecta todas las instancias de una clase con una funcionalidad adicional de delimitar instancias separadas de cualquier clase de segmento. Por lo tanto, también se conoce como la incorporación de la detección de objetos y la funcionalidad de segmentación semántica [4].

La segmentación por instancias tiene un formato de salida más enriquecedor, ya que crea un mapa

de segmentos para cada categoría e instancia de esa clase. En pocas palabras, considerando que se tiene una imagen con perros y gatos. Al ejecutar un modelo de segmentación por instancias en esa imagen, tú puedes localizar las bounding boxes de cada perro y gato, trazar mapas de segmentación para cada perro y gato, y así, contar cuantos perro y gatos hay en la imagen [5].

La segmentación de instancias implica identificar los límites de los objetos a nivel de pixel detallado, lo que convierte en una tarea compleja de realizar. Este modelo contiene 2 partes significantes:

- Detección de objetos
- Segmentación sentencia

Diferencias entre segmentación por instancias y semantica:

Tabla 1. Caracteristicas de segmentación semantica y por instacias.

Segementacion semantica	Segmentacion por instacnias
Para cada pixel de la imagen dada, detecta la categoria de objetos a lo que pertenece, donde todas las categorias/etiquetas son conocidas por el modelo	Para cas pixel de la imagen, identifica la instancia del objeto a la que pertenece. Se sumerge mas a profundidad que la segmentación semántica y diferencia dos objetos con la misma etiqueta
Ejemplo: No puede distinguie entre diferentes instancias de la misma categoría, es decir, todas las sillas están marcadas en azul	Ejemplo: Puede distinguir entre diferentes intancias de la misma categoria, es decir, diferente silla es distinguida por colores diferentes
En primer lugar, se detecta el objeto, y despues se etiqueta cada píxel	Se trata de un híbrido de anotación de detección de objetos y segmentación semántica
Lista de base de datos que es compatible: Stanford Background Dataset, Microsoft COCO Dataset, MSRC Dataset, KITTI Dataset, and Microsoft AirSim Dataset.	Lista de base de datos que es compatible: LiDAR Bonnetal Dataset, HRSID (High-Dimension SAR Images Dataset), SSDD (SAR Ship Detection Dataset), Pascal SBD Dataset, and iSAID (A Large Scale Aerial Images Dataset).

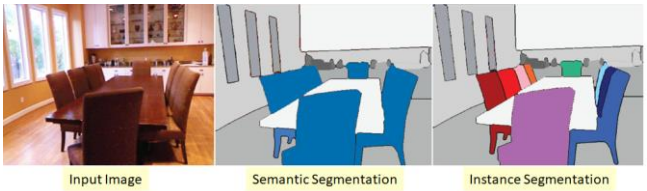

Figura 1. Segmentación semántica y por instancias.

Base de datos

La base de datos consta de imágenes con extensión JPEG, las cuales fueron extraídas de juegos de tenis en una transmisión televisiva, Se recolectaron 938 imágenes y se les aplico aumento de datos para el entrenamiento para sumar 3339 en total, Estas imágenes fueron segmentadas manualmente en 2 categorías: jugador de tenis y pelota de tenis, como se muestra en la Figura 2.

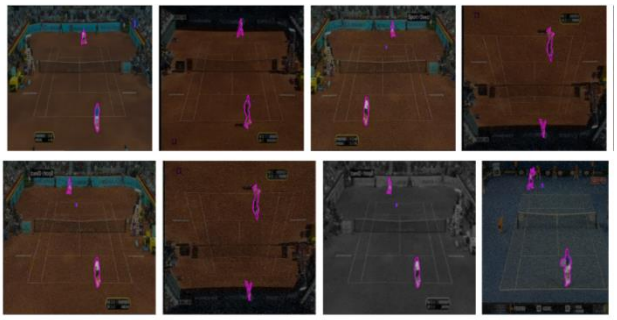

Figura 2. Base de datos propia de jugadores de partida de tenis.

II. MARCO TEÓRICO

Encoder-Decoder

Un encoder es una parte de una red neuronal que se encarga de procesar y comprimir la información de entrada. La compresión se realiza mediante la reducción de la dimensionalidad de la información, es decir, se eliminan algunas características de la información original. El objetivo es generar una representación compacta de la información que se puede utilizar para una tarea específica.

Por otro lado, un decoder es una parte de una red neuronal que se encarga de recuperar la información comprimida por el encoder. El objetivo es generar una representación detallada de la información original a partir de la representación comprimida. El decoder utiliza la información comprimida para reconstruir la información original lo más precisamente posible.

En conjunto, el encoder y el decoder forman una arquitectura de red neuronal llamada autoencoder,

que es utilizada para tareas de compresión y reconstrucción de información. Es común utilizar esta arquitectura para tareas de aprendizaje no supervisado, donde no se cuenta con etiquetas de entrenamiento (véase Figura 3) [6].

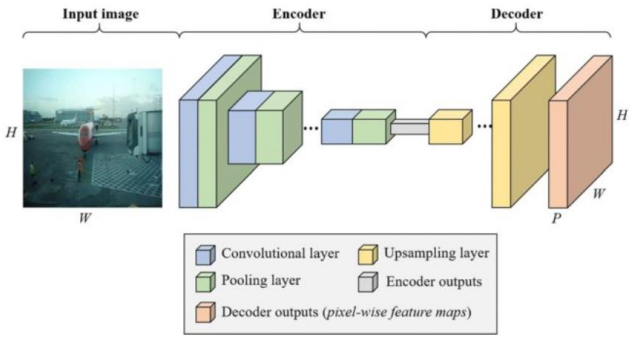

Figura 3. Estructura de un encoder-decoder.

Modelo FCN_8

FCN-8 es una arquitectura de red neuronal para tareas de segmentación de imágenes, desarrollada en 2015. Es una extensión de la arquitectura de red neuronal tradicional, Fully Convolutional Network (FCN).

La principal diferencia entre FCN-8 y las redes neuronales tradicionales es que FCN-8 tiene una estructura encoder-decoder, donde el encoder reduce la resolución espacial de la imagen a medida que se aumenta la profundidad de la red y el decoder aumenta la resolución de nuevo. FCN-8 utiliza técnicas de "upsampling" y "skip connections" para recuperar la información perdida en el proceso de compresión, permitiendo a la red generar una representación detallada de la imagen. FCN-8 ha sido utilizado con éxito en varias tareas de segmentación de imágenes, incluyendo la segmentación de objetos y la segmentación de múltiples clases. Es una arquitectura popular en la comunidad de investigación en visión por computadora [7].

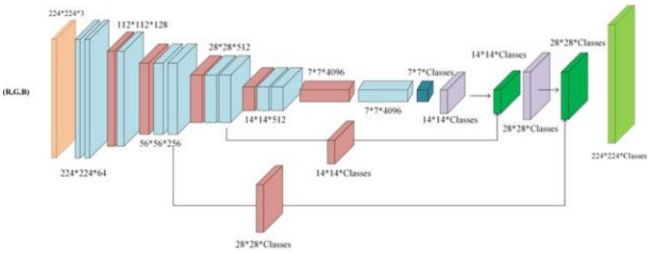

Figura 4. Estructura del modelo FCN_8.

Modelo VGG_16 segnet

VGG-16 es una arquitectura de red neuronal desarrollada para tareas de clasificación de imágenes. Fue presentado en 2014, es una versión más profunda de la arquitectura VGG. Es una red neuronal de tipo "Convolucional Neural Network" (CNN) basada en la idea de utilizar varias capas de filtros de 3x3 para extraer características de la imagen. Al utilizar varias capas de filtros, VGG-16 puede aprender características cada vez más complejas a medida que se profundiza en la red.

Aunque fue desarrollado originalmente para tareas de clasificación, VGG-16 también ha sido utilizado en tareas de segmentación de imágenes, mediante la adición de capas adicionales para generar máscaras de segmentación. Sin embargo, no es una arquitectura específica para la segmentación, sino que se ha utilizado como una base para desarrollar otras arquitecturas para la segmentación, ya que es una red neuronal muy potente y bien establecida en la comunidad de investigación en visión por computadora [8].

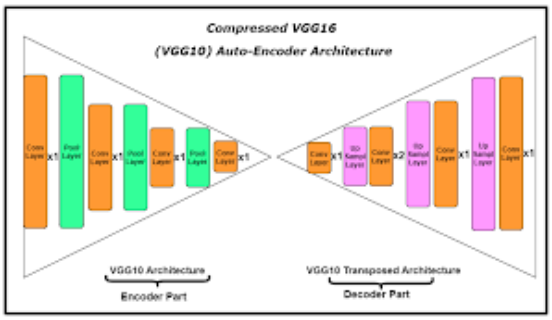

Figura 5. Estructura del modelo vgg_16.

Modelo VGG_Unet

VGG-Unet es una variante de la arquitectura U-Net, que combina las características de las redes

VGG con las de U-Net. Es una arquitectura de red neuronal desarrollada para tareas de segmentación de imágenes. La arquitectura VGG-Unet utiliza la estructura de encoder-decoder de U-Net, pero utiliza las capas de VGG como encoder en lugar de utilizar capas de convolución ordinarias. Esto permite a la red aprender características más complejas y detalladas de la imagen de entrada. Además, VGG-Unet utiliza conexiones "skip" entre las capas encoder y decoder para permitir la propagación de información detallada a través de todas las capas de la red. VGG-Unet ha demostrado tener buenos resultados en tareas de segmentación de imágenes, especialmente en tareas de segmentación de imágenes médicas y de satélite [9].

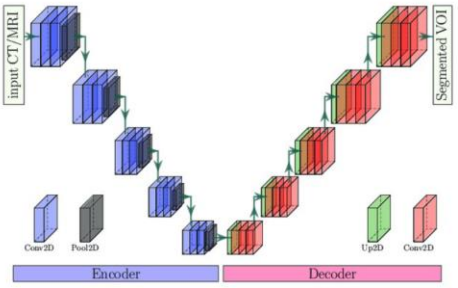

Figura 6. **Estructura del modelo VGG_Unet.**

Modelo YOLOV8

YOLOv8 es una arquitectura de red neuronal desarrollada para tareas de detección de objetos en imágenes. YOLO (You Only Look Once) es un enfoque de detección de objetos que consiste en dividir la imagen en pequeñas cajas y utilizar una red neuronal para clasificar cada caja como conteniendo o no un objeto específico.

YOLOv8 es una versión más reciente de la arquitectura YOLO, que ha sido mejorada en términos de precisión y velocidad de procesamiento. Aunque YOLOv8 fue desarrollado para tareas de detección de objetos, también se ha utilizado para la segmentación de imágenes, mediante la generación de máscaras de segmentación a partir de las cajas de detección. Sin embargo, YOLOv8 no es un modelo específico para la segmentación, sino que se ha utilizado como una base para desarrollar otras arquitecturas para la segmentación, ya que es una red neuronal muy potente y bien establecida en la comunidad de investigación en visión por computadora [10].

III. ENTRENAMIENTO DE MODELO

El modelo fue entrenado en los 4 modelos con las siguientes características:

Tabla 2. **Resultados del entrenamiento para cada modelo.**

Modelo	Tamaño de entrada	Épocas	Exactitud de entrenamiento
FCN_8	640x640	5	0.9955
VGG_16_segnet	640x640	5	0.9953
VGG_Unet	640x640	5	0.9965
YoloV8	600x600	50	0.4260

IV. EVALUACIÓN

La métrica de evaluación es la media IOU (intersección de la unión) y también IOU por clase. Obteniendo los siguientes resultados [11].

Modelo FCN_8

Este modelo consta de las siguientes características en 3 clases (fondo, bola, jugador de tenis):

Tabla 3. **Resultados de las métricas del modelo FCN_8.**

Media IU	IU por clase	
0.5270	Fondo	0.995
	Bola	0
	Jugador de tenis	0.5859

Figura 7. **Resultado de segmentación del modelo FCN_8 con 3 clases.**

Como se observa, no fue posible identificar la bola con este modelo, por lo que no resultó ser del todo certero para esta aplicación.

Modelo VGG_16_Segnet

Este modelo consta de las siguientes características en 3 clases (fondo, bola, jugador de tenis):

Tabla 4. Resultados de las métricas del modelo VGG_16_segnet.

Media IU	IU por clase	
0.5053	Fondo	0.9945
	Bola	0
	Jugador de tenis	0.5215

Figura 8. Resultado de segmentación del modelo VGG_16_Segnet con 3 clases.

Con este modelo tampoco fue posible identificar la bola, pues tiene un tamaño pequeño a pesar de las 3 clases definidas inicialmente.

Modelo VGG_Unet

Este modelo consta de las siguientes características en 3 clases (fondo, bola, jugador de tenis):

Tabla 5. Resultados de las métricas del modelo VGG_Unet.

Media IU	IU por clase	
0.5784	Fondo	0.9955
	Bola	0.13085
	Jugador de tenis	0.60

Figura 9. Resultado de segmentación del modelo VGG_unet con 3 clases.

Con este modelo fue posible obtener un IU mayor a 0 en la bola, por lo que es posible distinguirla en algunos casos, aunque no sea muy precisa en realizarlo.

Modelo basado en YOLOV8

En este caso las métricas basadas en detección son utilizadas para conocer la eficacia de la red.

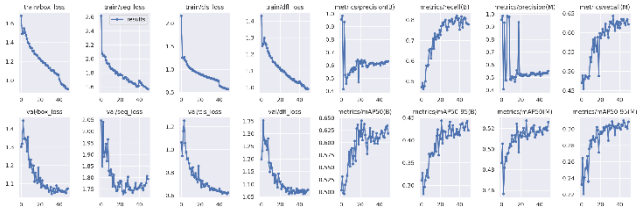

Figura 10. Métricas para el modelo basado en YOLOV8.

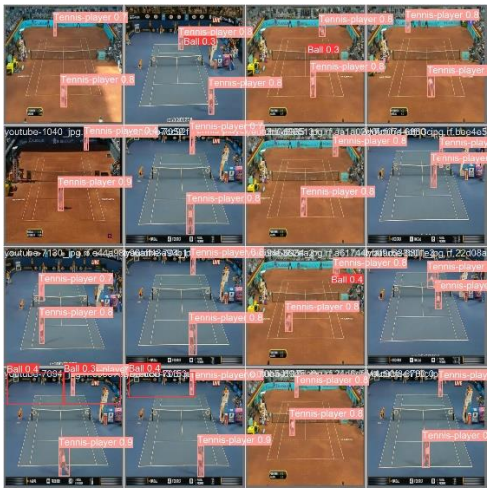

Figura 11. Resultado de segmentación del modelo YOLOV8 con 2 clases.

En este modelo solo se tienen dos clases que son la bola y los jugadores. Sin embargo, en la clase de la bola no es posible distinguir de manera correcta,

pues su probabilidad fue muy baja en un área muy grande de la detección del cuadro.

VI. CONCLUSIÓN

Se pudo observar el funcionamiento de un modelo de segmentación semántico y de instancias, en donde principalmente vario el tipo de datos de entrada en el formato requerido para poder procesar la información.

Posteriormente fue posible observar cómo es que los modelos les costaban más trabajo poder segmentar objetos tan pequeños, se cree que puede ser por el tamaño de la imagen al bajar la resolución a 640x640. Además de que falto incrementar la cantidad de épocas para entrenar, aunque la capacidad de computo no sea la suficiente para realizar esta tarea de manera veloz.

El modelo basado en Unet fue el que obtuvo los mejores resultados, esto se debe a que es el modelo mas especializado para este tipo de problemas de segmentación siendo este aplicado en segmentación semantica. Tambien se considera que la cantidad de imágenes debió aumentar para mejorar la precisión así como tambien poner otros ambientes y situaciones del juego en los que los jugadores tal vez no se aprecian tan bien o se encuentran en posiciones atípicas del juego.

REFERENCIAS

[1] C. Silva, "Modelamiento Semántico del Entorno de un Robot utilizando información RGB-D", 2016. doi: 10.13140/RG.2.2.28418.17609.

[2] O. A. Soto-Orozco, A. D. Corral-Sáenz, C. E. Rojo-González, y J. A. Ramírez-Quintana, "Análisis del desempeño de redes neuronales profundas para segmentación semántica en hardware limitado", *ReCIBE. Revista electrónica de Computación, Informática, Biomédica y Electrónica*, vol. 8, núm. 2, pp. 1–21, 2019.

[3] D. Iglesias, "Segmentación de Imágenes con Redes Convolucionales", el 18 de abril de 2021. https://www.iartificial.net/segmentacion-imagenes-redes-convolucionales/ (consultado el 18 de enero de 2023).

[4] A. Gómez, F. León-Pérez, M. Plazas-Wadynski, y F. Martínez-Carrilo, "Segmentación multinivel de patrones de Gleason usando representaciones convolucionales en imágenes histopatológicas", *TecnoLógicas*, vol. 24, p. e2132, dic. 2021, doi: 10.22430/22565337.2132.

[5] D. M. Woerdemann, "Segmentación de instancias simplificada de células y núcleos mediante el aprendizaje profundo". https://www.olympus-lifescience.com/es/discovery/instance-segmentation-of-cells-and-nuclei-made-simple-using-deep-learning/ (consultado el 18 de enero de 2023).

[6] A. Dertat, "Applied Deep Learning - Part 3: Autoencoders", *Medium*, el 8 de octubre de 2017. https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798 (consultado el 18 de enero de 2023).

[7] S. Piramanayagam, E. Saber, W. Schwartzkopf, y F. Koehler, "Supervised Classification of Multisensor Remotely Sensed Images Using a Deep Learning Framework", *Remote Sensing*, vol. 10, p. 1429, sep. 2018, doi: 10.3390/rs10091429.

[8] A. Kebir, M. Taibi, y F. Serradilla, "Compressed VGG16 Auto-Encoder for Road Segmentation from Aerial Images with Few Data Training".

[9] T. Ghosh, Md. K. Hasan, S. Roy, Md. A. Alam, E. Hossain, y M. Ahmad, *Multi-class probabilistic atlas-based whole heart segmentation method in cardiac CT and MRI*. 2021.

[10] "Segmentation - Ultralytics YOLOv8 Docs". https://docs.ultralytics.com/tasks/segmentation/ (consultado el 17 de enero de 2023).

[11] "Evaluación de calidad en la segmentación de imágenes - PDF Descargar libre". https://docplayer.es/10131855-Evaluacion-de-calidad-en-la-segmentacion-de-imagenes.html (consultado el 17 de enero de 2023).