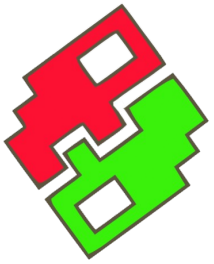
Fundación Tarpuy

Segundo Nivel Ingenia

Segundo semestre 2023



Redes Neuronales

Aplicada a la detección de señales de tránsito

Tutor: Leandro Borgnino

Integrantes: Medina Santiago, Esteban Suarez y Alfonso Mouton

Índice

Introducción

1.1. Motivación

En el vertiginoso avance de la tecnología, la aplicación de redes neuronales en el procesamiento de imágenes ha emergido como un catalizador revolucionario en diversas disciplinas. Entre las múltiples facetas que abarca esta amalgama de inteligencia artificial y visión computarizada, la detección y clasificación de señales de tránsito se destaca como un campo de estudio de gran relevancia e impacto práctico. La seguridad vial es una preocupación global de suma importancia, y el tráfico vehicular se presenta como un escenario dinámico y complejo donde la correcta interpretación de señales juega un papel crucial. La detección automatizada y la clasificación precisa de señales de tránsito no solo pueden potenciar la eficiencia de los sistemas de transporte, sino que también desempeñan un papel esencial en la prevención de accidentes y la mejora de la movilidad urbana.

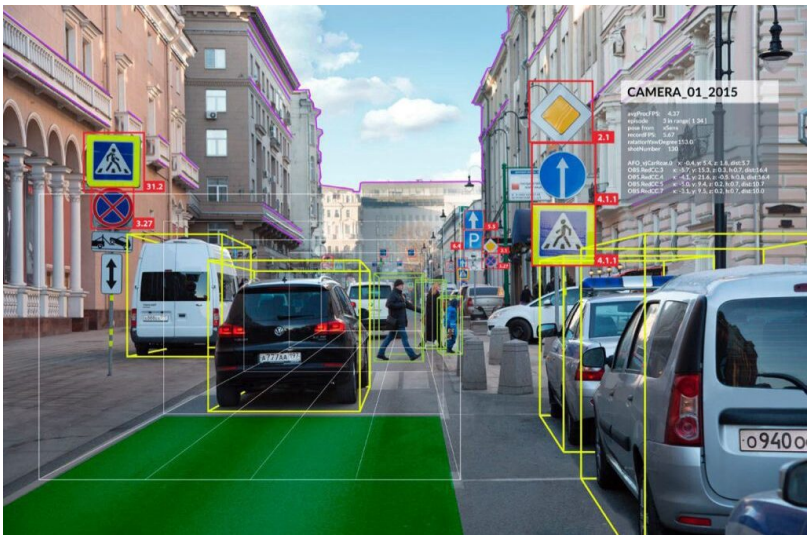


Figura 1: Reconocimiento de Objetos.

El crecimiento exponencial de datos visuales en entornos urbanos y la necesidad de respuestas rápidas ante señales cambiantes hacen imperativa la adopción de enfoques avanzados. Acá es donde las redes neuronales toman importancia, porque su capacidad para aprender patrones complejos a partir de grandes conjuntos de datos permite el desarrollo de modelos capaces de discernir con precisión las señales de tráfico en imágenes, incluso en condiciones adversas. En este contexto, la motivación subyacente es impulsar la aplicación de redes neuronales en el ámbito específico de la detección y clasificación de señales de tránsito.

1.2. Objetivos

Los objetivos del proyecto son los siguientes:

1. **Pre-procesamiento de los datos:** Se llevará a cabo un pre-proceso en los datasets que tendrán relevancia en el entrenamiento de los distintos modelos.
2. **Implementación de Arquitecturas Neuronales:**
   a) **Desarrollo del modelo VGG-16:** Desarrollar la arquitectura de la red neuronal VGG-16 para la obtención de mapas de características, quitando capas para obtener resultados que necesitaremos para la siguiente etapa.
   b) **Desarrollo del modelo ROI-Pooling:**
   - Diseñar la arquitectura de la red neuronal ROI Pooling para identificar la presencia de señales de tránsito en los mapas de características obtenidos por VGG-16.

- Entrenar la red ROI Pooling con diversos conjuntos de mapas de características generados a partir de imágenes de tráfico.

c) **Implementación del modelo de clasificación:**

- Investigación y evaluación de diferentes arquitecturas de redes neuronales. Se seleccionará la arquitectura más adecuada para el clasificador de señales de tránsito, teniendo en cuenta la precisión, la eficacia computacional y el tamaño del modelo.
- Entrenar la red de clasificación utilizando un conjunto de datos exclusivo que incluye imágenes de señales de tránsito etiquetadas.

3. **Integración de las redes neuronales:** Integrar las arquitecturas VGG-16, ROI Pooling y la red de clasificación para formar un modelo coherente y funcional de clasificación de señales de tránsito.

4. **Evaluación del rendimiento del modelo:** Evaluar la precisión y eficacia del modelo completo mediante métricas relevantes de clasificación.

5. **Comunicación:** Analizar críticamente los resultados, identificar posibles mejoras y proporcionar recomendaciones para futuras investigaciones.

Marco Teórico

2.1. Detección de Objetos

La detección de objetos constituye una rama esencial en el campo del procesamiento de imágenes y la visión por computadora. Su objetivo principal es discernir y localizar la presencia de uno o más objetos dentro de una imagen completa, asignándoles una identidad específica. Las técnicas de detección de objetos se han desarrollado de manera significativa en respuesta a la creciente necesidad de sistemas capaces de comprender y responder a entornos visuales complejos. Uno de los enfoques más destacados y exitosos en este ámbito es el uso de redes neuronales convolucionales (CNN), que han demostrado una eficacia excepcional en la extracción de características relevantes de las imágenes. La arquitectura típica para la detección de objetos a menudo involucra dos etapas cruciales: la generación de propuestas y la clasificación de esas propuestas. En la primera etapa, se utilizan técnicas como Region Proposal Networks (RPN) para proponer regiones de interés que podrían contener objetos. Posteriormente, estas regiones son clasificadas y refinadas utilizando capas de clasificación y regresión.

Figura 2: Reconocimiento facial

Un enfoque más reciente y potente en la detección de objetos es la utilización de modelos de detección de objetos de una sola etapa, como YOLO (You Only Look Once) y SSD (Single Shot Multibox Detector). Estos modelos permiten la detección de objetos en tiempo real al abordar la tarea de manera conjunta, prediciendo las clases y las ubicaciones de los objetos de una sola vez. Además, existen distintos tipos de detecciones y clasificaciones:

- Clasificación de la imagen.
- Clasificación con Localización.
- Detección.

2.2. Clasificación de la imágen

En líneas generales, la clasificación de imágenes implica el uso de CNN seguidas de capas totalmente conectadas. Estas redes convolucionales son particularmente eficientes en la extracción de características relevantes, como bordes, texturas y patrones, mientras que las capas totalmente conectadas permiten la interpretación global de estas características para realizar la clasificación final. Por ejemplo, en el contexto específico de la detección de vehículos, el objetivo de la clasificación se centraría en determinar la probabilidad de que la imagen contenga un automóvil. Esta probabilidad se obtiene al alimentar la imagen a través de la red neuronal, que ha sido previamente entrenada para reconocer patrones asociados con la presencia de vehículos.

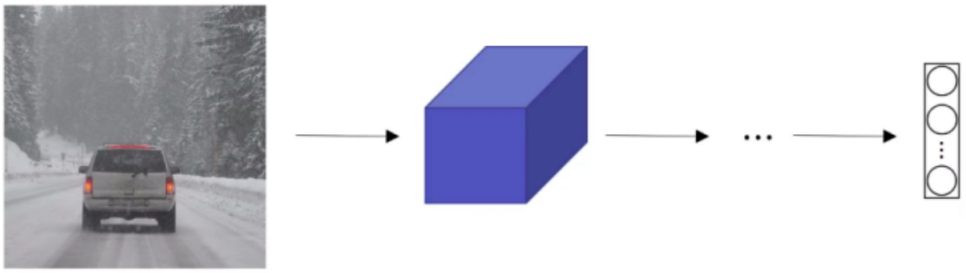


Figura 3: Representación del proceso de clasificación

2.3. Clasificación con Localización

Para refinar aún más esta tarea, se pueden agregar neuronas de salida adicionales que proporcionen información sobre la localización del vehículo. Este enfoque, a menudo utilizado en sistemas de detección y clasificación de objetos, utiliza una bounding box (caja delimitadora) para representar la región en la que se encuentra el objeto de interés. En el caso de la detección de vehículos, esta `bounding box` se define mediante cuatro valores: **bx** (coordenada x del centro), **by** (coordenada y del centro), **bh** (altura) y **bw** (ancho)

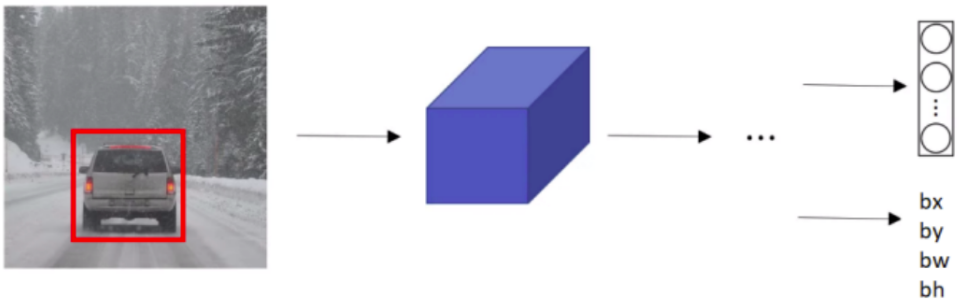


Figura 4: Representación del proceso de clasificación con localización

La inclusión de esta información espacial permite no solo identificar la presencia de un vehículo en la imagen, sino también delimitar su ubicación precisa. Este enfoque, combinado con técnicas avanzadas de clasificación, potencia la capacidad del sistema para comprender la escena visual en su totalidad y, en el caso específico de la detección de vehículos, proporcionar información detallada sobre su posición en la imagen.

Por ejemplo, la salida de la red de clasificación con localización, considerando tres clases de objetos diferentes, sería la siguiente:

$$y = [p_c \ b_x \ b_y \ b_w \ c_1 \ c_2 \ c_3]^T$$

Si ahora planteamos la función de pérdida de la salida (con MSE) tenemos:

- Si $p_c = 1$

$$L(y, \hat{y}) = (\hat{y_1} - y_1)^2 + (\hat{y_2} - y_2)^2 + \ldots + (\hat{y_8} - y_8)^2$$

- Si $p_c = 0$

$$L(y, \hat{y}) = (\hat{y_1} - y_1)^2$$

Por lo general se utiliza la función de pérdida log likelihood para las clases, mse para las coordenadas de la región limitante y logistic regression para p_c.

2.4. Redes STN

Las Redes STN, o Redes de Transformadores Espaciales (Spatial Transformer Networks), representan un enfoque innovador en el campo de las Redes Neuronales Convolucionales (CNN). Estas redes han sido diseñadas para abordar la invarianza espacial en las entradas, permitiendo que la red aprenda a manejar variaciones en rotación, traslación y escala de manera más efectiva.

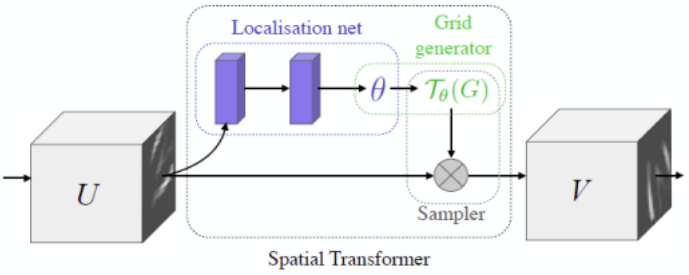


Figura 5: Representación de un Transformador espacial

El núcleo distintivo de una Red STN radica en su incorporación de módulos de "transformadores espaciales". Estos módulos proporcionan a la red la capacidad de realizar transformaciones geométricas sobre las imágenes de entrada de manera automática durante el proceso de aprendizaje. Esta capacidad es fundamental para mejorar la robustez y la generalización del modelo, ya que permite que la red se vuelva espacialmente invariante a las variaciones mencionadas. Las transformaciones espaciales, tales como rotaciones y traslaciones, pueden introducir variaciones significativas en las características visuales de una imagen. Por ejemplo, en el contexto de la detección de objetos en imágenes de tráfico, la rotación de una señal de tránsito o su traslación en la imagen pueden desafiar la capacidad de una red convencional para reconocerla de manera efectiva.

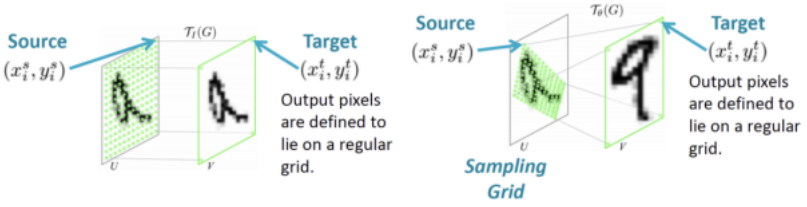


Figura 6: Representación de la aplicación de una transformación.

Los módulos de transformadores espaciales en una Red STN actúan como mecanismos de atención, permitiendo a la red aprender a enfocarse en regiones específicas de la entrada y realizar ajustes geométricos según sea necesario. Esta adaptabilidad a las variaciones espaciales mejora la capacidad del modelo para capturar patrones relevantes independientemente de la orientación o posición en la imagen.

Ahora bien, para el caso de una red de localización, donde se tiene un mapa de características de entrada U, con canales de ancho W y alto H, las salidas son y los parámetros de transformación T, existiendo distintos tipos de transformaciones posibles. Una de esas transformaciones es la Transformación afín, donde dependiendo de los valores en la matriz, podemos transformar (X_1, Y_1) a (X_2, Y_2) con diferentes efectos.

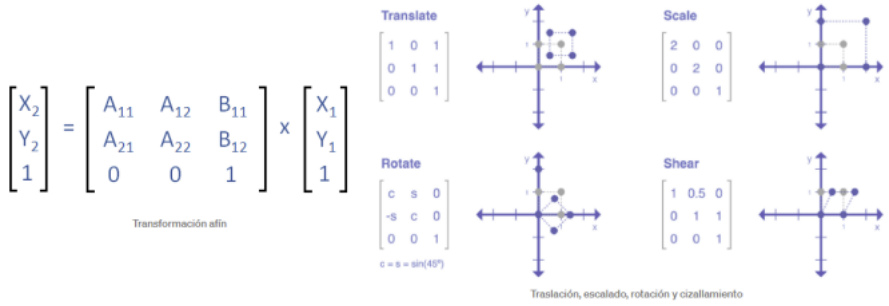


Figura 7: Representación de algunos efectos

2.5. Etapa de la detección de objetos

En el campo de la detección de objetos, las arquitecturas aplicadas a menudo se estructuran en dos etapas distintas, cada una desempeñando un papel crucial en el proceso global de reconocimiento y localización de objetos en imágenes. Estas etapas son la detección de la región y la detección y clasificación del objeto.

1. Detección de la Región En la primera etapa, la detección de la región se centra en identificar áreas candidatas que podrían contener objetos de interés. Diversas técnicas han sido desarrolladas para esta tarea, entre las cuales se destacan la ventana deslizante y la búsqueda selectiva, entre otras similares. La ventana deslizante implica el escaneo sistemático de la imagen mediante una ventana móvil, evaluando cada región para determinar la probabilidad de contener un objeto. Por otro lado, la búsqueda selectiva utiliza propuestas generadas previamente para enfocarse en áreas más prometedoras de la imagen, reduciendo así la carga computacional.

2. Detección y Clasificación del Objeto: Una vez identificadas las regiones de interés, la siguiente etapa involucra la detección y clasificación del objeto dentro de estas regiones. Aquí, se pueden utilizar clasificadores clásicos o redes neuronales convolucionales entrenadas específicamente para reconocer patrones visuales asociados con categorías de objetos. Este paso es esencial para asignar una etiqueta a cada objeto detectado, proporcionando información sobre su naturaleza y características

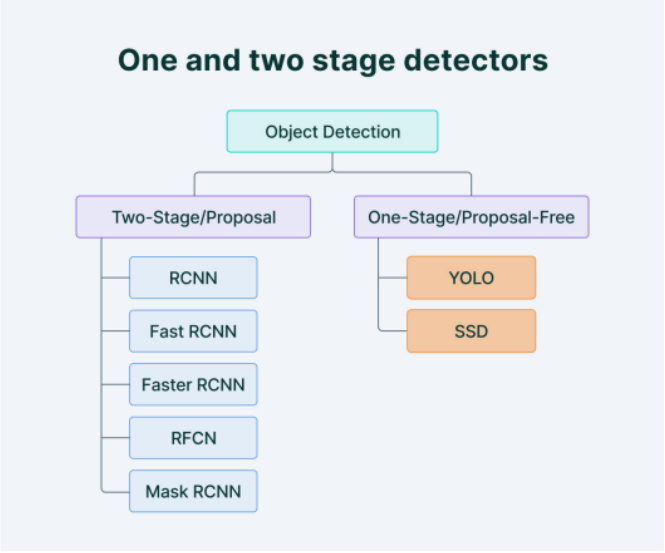


Figura 8: Esquema de detectores por etapas

2.5.1. Arquitecturas Integradas

Aunque las dos etapas mencionadas anteriormente son comunes, han surgido arquitecturas más avanzadas que realizan ambas tareas simultáneamente. Estas arquitecturas integran la detección de la región y la clasificación del objeto en una única red neuronal, permitiendo una inferencia más eficiente y rápida. En el panorama de la detección de objetos, la elección entre arquitecturas de una o dos etapas depende de los requisitos específicos de la aplicación y las limitaciones computacionales.

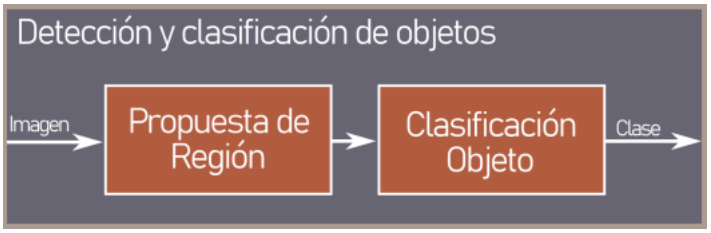


Figura 9: Arquitectura con dos etapas

2.6. Ventana deslizante

La técnica de ventana deslizante, que implica el desplazamiento de rectángulos a través de una imagen en busca de objetos, se convierte en una estrategia aún más potente cuando se incorporan optimizaciones clave. Este enfoque exhaustivo se beneficia de la flexibilidad para cambiar el tamaño de la imagen o de la propia ventana deslizante, permitiendo la obtención de cajas de delimitaciones más precisas.

Figura 10: Representación de la técnica

Cuando se trabaja con imágenes de tamaños variados, ajustar la escala de la ventana deslizante es esencial para garantizar la detección efectiva de objetos en diferentes contextos. Posteriormente, basándose en las ventanas donde se detecta el objeto en imágenes más pequeñas, se puede escalar nuevamente y unir las detecciones. Este proceso contribuye significativamente a obtener resultados más precisos y detallados, especialmente en escenarios donde la escala de los objetos varía considerablemente.

Figura 11: Representación de la técnica

La versatilidad de la ventana deslizante también puede conducir a situaciones en las que varias cajas de delimitaciones detectan el mismo objeto. El desafío radica en seleccionar la mejor candidata entre estas detecciones redundantes. Para abordar este problema, se emplea la técnica de **Non-Maximum Suppression** (supresión de no máximos) entre las candidatas. Esta estrategia se centra en retener únicamente la detección más confiable, descartando las demás para evitar duplicaciones innecesarias.

Figura 12: Representación de la técnica NMS

La evaluación de la calidad de las detecciones se realiza mediante la métrica Intersection Over Union (Intersección sobre Unión, IoU). Esta métrica calcula la proporción entre la intersección y la unión de dos regiones delimitadas por cajas. Al establecer un umbral específico de IoU, se puede determinar la superposición aceptable entre dos cajas para considerarlas como duplicadas o no. Esto asegura que la mejor candidata seleccionada sea la que mejor se ajusta a la verdadera posición y forma del objeto.

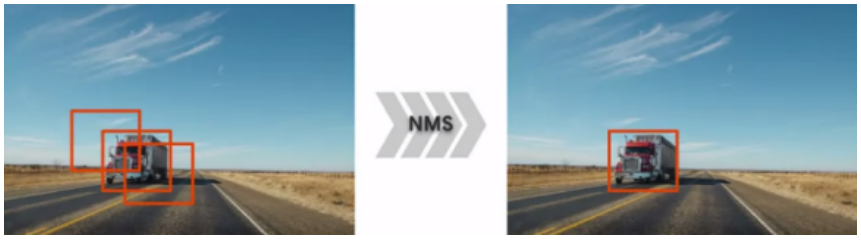


Figura 13: Representación de la técnica NMS

2.6.1. Otros datos

La técnica de ventana deslizante, a pesar de ser efectiva en la detección de objetos, enfrenta desafíos computacionales significativos, especialmente en términos de eficiencia. Su alto costo computacional se debe a que cada recorte generado por el desplazamiento de la ventana debe procesarse individualmente por la red convolucional, siendo más intensivo en recursos con imágenes de alta resolución o detecciones más precisas.

La situación empeora al buscar un desplazamiento más preciso, ya que explorar un espacio de búsqueda más fino implica procesar más regiones, aumentando exponencialmente la carga computacional. La ventana deslizante no es óptima para Redes Neuronales Convolucionales (CNN), ya que la complejidad de las CNN hace que el procesamiento independiente de cada recorte sea ineficiente y costoso.

Para superar estas limitaciones, se propone la ventana deslizante convolucional, que optimiza el proceso al introducir la convolución directamente en la ventana. Esto permite a la red compartir cálculos entre regiones superpuestas, reduciendo redundancias y mejorando la eficiencia global del modelo.

2.7. Intersección sobre la Union (IoU)

La métrica de Intersección sobre la Unión (IoU) es como un elemento crucial en la evaluación de la precisión y calidad de las predicciones en la detección de objetos. Esta métrica proporciona una medida del grado de superposición entre dos regiones delimitantes, ofreciendo una indicación clara de la similitud y, por ende, de la efectividad del modelo predictivo.

El cálculo de IoU está dado por la siguiente expresión:

$$IoU = \frac{\text{área de la intersección}}{\text{área de la unión}}$$

Donde el numerador representa el área compartida entre las dos regiones delimitantes, y el denominador representa el área total cubierta por ambas. Este cálculo proporciona un valor normalizado que varía entre 0 y 1, donde 0 indica ninguna superposición y 1 indica una coincidencia perfecta entre las regiones.

La interpretación del resultado de IoU es directa:

- A medida que el valor de IoU se acerca a 1, se indica una mayor similitud y precisión en la predicción
- Un IoU de 0.5 se considera un umbral comúnmente aceptado para determinar si una predicción es correcta
- Si el IoU es mayor a 0.5, se considera que la predicción es precisa, lo que implica que la región predicha se superpone significativamente con la región real del objeto.

Esta métrica es especialmente valiosa en situaciones donde es esencial evaluar no solo la detección de un objeto, sino también la precisión de su ubicación y forma predicha. Al establecer un umbral significativo, se puede establecer un estándar para la aceptabilidad de las predicciones, contribuyendo así a la toma de decisiones en la optimización de modelos de detección de objetos.

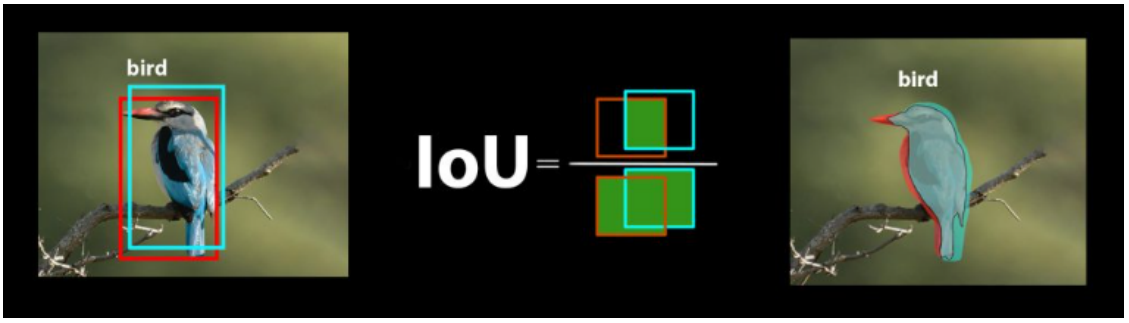


Figura 14: Representación de la métrica IoU

2.8. Non-Max Supression

La técnica de Non-Maximum Suppression (NMS) es una estrategia esencial en la postprocesamiento de detecciones, particularmente cuando se enfrenta a la presencia de múltiples celdas o ventanas que comprenden un solo objeto. El objetivo principal de NMS es seleccionar la ventana que mejor encuadre al objeto en cuestión, eliminando redundancias y asegurando una salida precisa. El proceso de Non-Maximum Suppression inicia evaluando las probabilidades asociadas con cada detección (pc) y seleccionando la ventana con la probabilidad más alta. Este paso inicial garantiza que la predicción más confiable se mantenga, sirviendo como referencia para la supresión de detecciones redundantes.

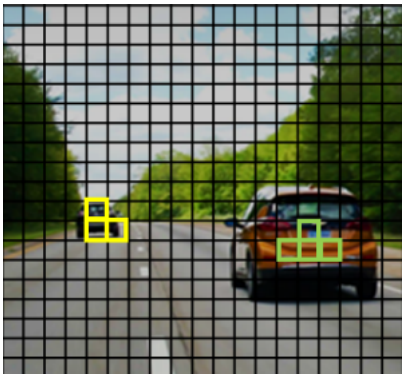


Figura 15: Referencia a la técnica NMS

A continuación, se examinan los rectángulos **cercanos** al rectángulo con la mayor probabilidad:

Se calcula el valor de solapamiento (IoU) entre estos rectángulos y el rectángulo de mayor probabilidad. Aquellos rectángulos que tienen un valor de IoU significativo con el rectángulo de referencia son suprimidos, ya que representan detecciones superpuestas o redundantes. Este proceso, además de supresión, garantiza que

solo se retenga la detección más confiable y precisa, eliminando aquellas que no aportan información adicional significativa. La técnica de Non-Maximum Suppression, por lo tanto, contribuye a la generación de resultados más limpios y coherentes en el contexto de la detección de objetos. Es importante destacar que, en escenarios donde hay varios objetos a detectar, NMS puede ejecutarse de manera independiente para cada salida, permitiendo así un manejo eficiente de múltiples detecciones en una única imagen, y así asegurando que la selección de ventanas se realice de manera óptima para cada objeto individual.

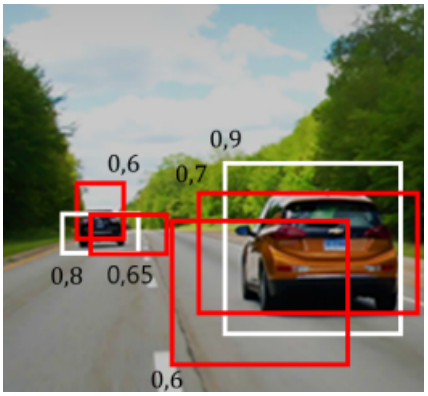


Figura 16: Representación de NMS+IoU

2.9. Mapa de densidad

Otra estrategia que se utiliza ampliamente es la de los mapas de densidad. La metodología de mapas de densidad comienza con la creación de una matriz de ceros del mismo tamaño que la imagen original. Por cada ventana cuyo puntaje (score) supere un umbral predeterminado, se incrementa en uno la cantidad en las posiciones que la ventana abarca en la matriz. Este enfoque tiene el propósito de construir un mapa que refleje la densidad de detecciones en diferentes regiones de la imagen.

Posteriormente, el proceso implica recortar las zonas de la matriz cuyos valores superan un umbral específico. Este paso refina la información, destacando las regiones con una densidad significativa de detecciones y descartando áreas con puntuaciones inferiores.

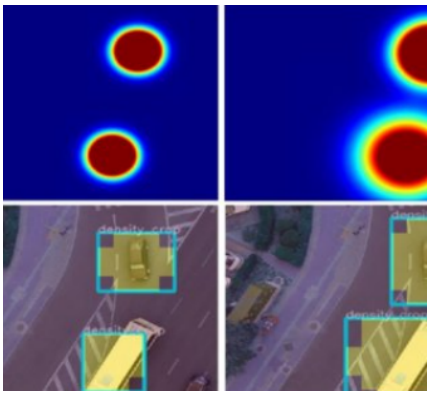


Figura 17: Ejemplo de un mapa de densidad

La utilización de mapas de densidad aporta varios beneficios al proceso de detección de objetos. Al construir una representación más rica de la distribución espacial de los objetos detectados, esta técnica puede ser especialmente valiosa en situaciones donde se busca comprender la concentración y la disposición precisa de los elementos de interés.

Es importante tener en cuenta que la elección de los umbrales juega un papel crucial en la efectividad de los mapas de densidad. Ajustar adecuadamente estos umbrales permite adaptar la técnica a las características específicas de la aplicación y optimizar la detección de objetos en diferentes contextos.

2.10. Ventana deslizante convolucional

La Ventana Deslizante Convolucional es una evolución de la técnica original de ventana deslizante en el ámbito de la detección de objetos. Esta adaptación introduce la convolución directamente en el proceso de exploración de la imagen, superando las limitaciones computacionales asociadas con la versión clásica.

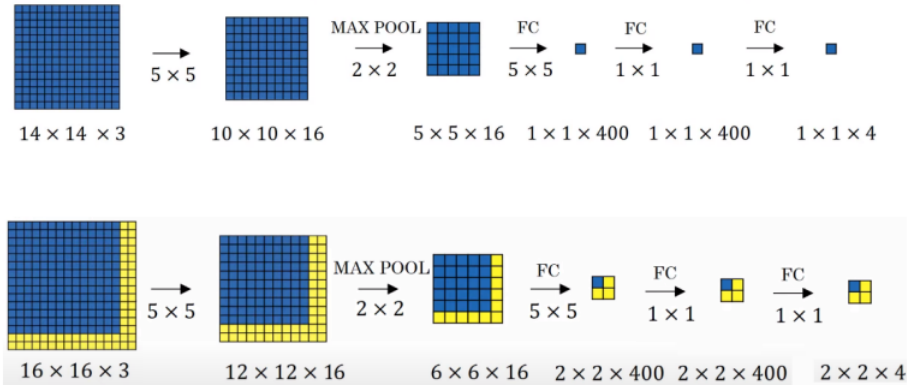


Figura 18: Ejemplo de funcionamiento

En la implementación convolucional de la ventana deslizante, se aprovecha la capacidad de las redes neuronales convolucionales (CNN) para realizar operaciones de convolución de manera eficiente. En lugar de procesar cada recorte de la imagen de manera independiente, la red convolucional comparte cálculos entre regiones superpuestas, reduciendo así la redundancia y optimizando el costo computacional asociado con la técnica de ventana deslizante tradicional.

El proceso se inicia con la generación de múltiples niveles de representaciones a través de capas convolucionales. Estas representaciones capturan gradualmente características de diferentes niveles de abstracción en la imagen. Luego, se desliza una ventana convolucional a través de estas representaciones, permitiendo que la red focalice su atención en regiones específicas y aprenda patrones jerárquicos de manera eficaz.

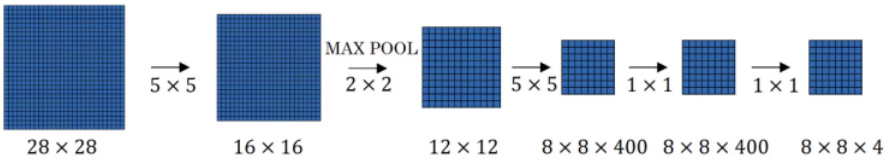


Figura 19: Proceso

La Ventana Deslizante Convolucional aborda la desventaja computacional de la ventana deslizante clásica al mejorar la eficiencia y la capacidad de aprendizaje de la red. Al utilizar capas convolucionales, se logra una mayor capacidad de generalización, lo que permite al modelo detectar objetos en diferentes escalas, orientaciones y contextos visuales.

Figura 20: Representación de la técnica

Además, esta técnica contribuye a la generación de mapas de características que resaltan áreas relevantes de la imagen. Estos mapas, combinados con capas de clasificación y regresión, permiten la detección precisa

y la asignación de categorías a los objetos identificados.

2.11. Predicción del cuadro delimitador

A pesar de la optimización que proporciona la implementación de la técnica de ventana deslizante en conjunto con redes neuronales convolucionales (CNN), surge una limitación crucial en la precisión de los cuadros delimitadores asociados a los objetos detectados.

Esta limitación se manifiesta en situaciones donde ninguna ventana deslizante coincide de manera precisa con la posición real del objeto, como un vehículo. Además, la forma y tamaño predeterminados de la ventana deslizante pueden no ser los más apropiados para delimitar con precisión la región de interés.

La falta de precisión en los límites del cuadro delimitador puede comprometer la exactitud global del sistema de detección de objetos. Esto se convierte en un desafío importante, ya que la correcta delimitación de la región ocupada por el objeto es esencial para comprender su ubicación y forma con precisión. Para abordar esta limitación, se exploran enfoques avanzados que buscan mejorar la predicción de los cuadros delimitadores, permitiendo una detección más precisa y detallada de los objetos en la imagen. Estos enfoques a menudo involucran técnicas de regresión que buscan ajustar dinámicamente los límites del cuadro en función de las características específicas del objeto detectado.

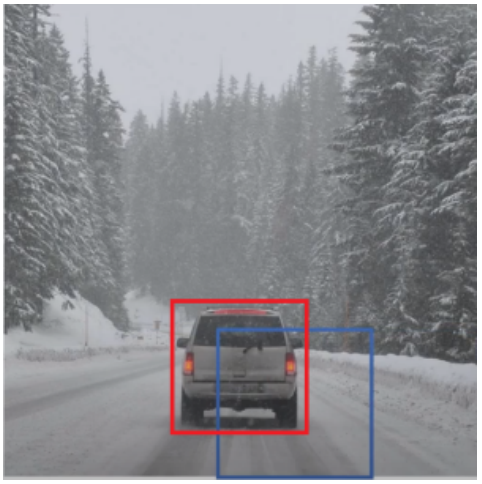


Figura 21: Ejemplo de predicción

2.12. R-CNN

El enfoque R-CNN es una metodología que incorpora redes neuronales convolucionales (CNN) basadas en regiones para implementar la búsqueda selectiva de objetos en una imagen. El proceso R-CNN inicia con la fase de búsqueda selectiva, donde se exploran alrededor de 2000 posibles regiones de interés en la imagen. Esta etapa utiliza la técnica de búsqueda selectiva para identificar áreas prometedoras que podrían contener objetos de interés. Esta primera fase de selección de regiones es esencial para reducir la complejidad computacional y centrar el análisis en áreas relevantes.

Posteriormente, cada región seleccionada se somete a un proceso de extracción de características utilizando una red neuronal pre-entrenada. Esta red, habitualmente diseñada para tareas de clasificación de imágenes a gran escala, se adapta para capturar las características específicas de las regiones de interés. Esta adaptación permite que la red aprenda representaciones significativas de los objetos contenidos en las regiones, contribuyendo a una detección más precisa.

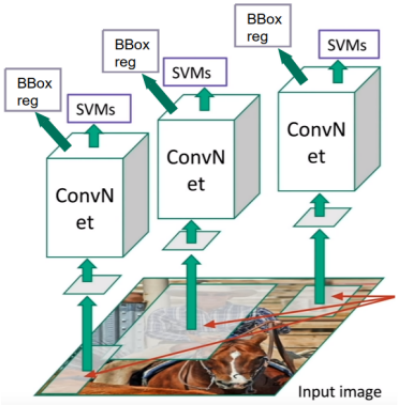


Figura 22: Representación de R-CNN

2.13. Fast R-CNN

Para superar las limitaciones de velocidad inherentes a la arquitectura R-CNN, surge Fast R-CNN. R-CNN, aunque efectiva, enfrenta desafíos computacionales significativos debido a su técnica de búsqueda selectiva en toda la imagen y al procesamiento lento de 2000 áreas de interés a través de las CNN. Fast R-CNN aborda estas limitaciones implementando una estrategia más eficiente. En lugar de enviar cada región de interés por separado a la red neuronal, Fast R-CNN integra una CNN que opera en toda la imagen.

Esta CNN de toda la imagen extrae características generales y crea feature maps que luego se utilizan para cada región de interés. Esta arquitectura optimizada ahorra tiempo y recursos computacionales al evitar el procesamiento repetitivo de la imagen completa para cada región. Al utilizar feature maps compartidos, Fast R-CNN logra una mayor eficiencia en la extracción de características, permitiendo una detección más rápida y precisa de objetos en la imagen.

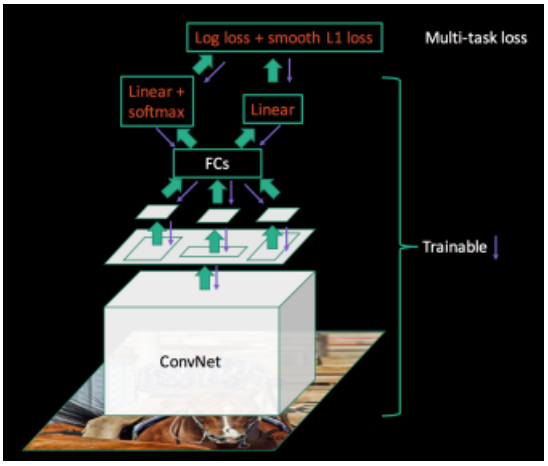


Figura 23: Representación de Fast R-CNN

2.14. Faster R-CNN

Faster R-CNN integra de manera eficiente el algoritmo de proposición de regiones de interés directamente en la red neuronal convolucional (CNN).

La principal innovación de Faster R-CNN radica en su capacidad para unificar el enfoque eficiente de extracción de características de Fast R-CNN con un algoritmo de proposición de regiones, creando así una arquitectura integral y ágil. Esta integración permite que la red proponga automáticamente las regiones de interés, eliminando la necesidad de procesos separados y mejorando drásticamente la eficiencia del modelo.

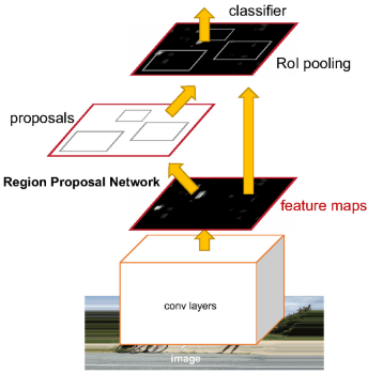


Figura 24: Representación de Faster R-CNN

Comparado con su predecesor, R-CNN, Faster R-CNN se destaca por su velocidad. Es 250 veces más rápido que R-CNN, y supera significativamente a Fast R-CNN siendo 25 veces más rápido. Estas mejoras en velocidad no solo aceleran el proceso de detección, sino que también abren posibilidades para la implementación de sistemas en tiempo real y aplicaciones de visión por computadora más exigentes.

2.15. YOLO(You Only Look Once)

2.15.1. Algoritmo

YOLO lleva a cabo la detección de objetos realizando la clasificación y localización en un solo paso. Su enfoque consiste en dividir la imagen en una cuadrícula SxS y prever, para cada elemento de la cuadrícula, N áreas de detección con un porcentaje de confianza asociado.

En total, YOLO predice SxSxN áreas, pero aplica un filtro, descartando aquellas con baja confidencia. Esta estrategia de filtrado contribuye a la generación de predicciones más precisas y confiables, centrándose en áreas con una alta probabilidad de contener objetos de interés.

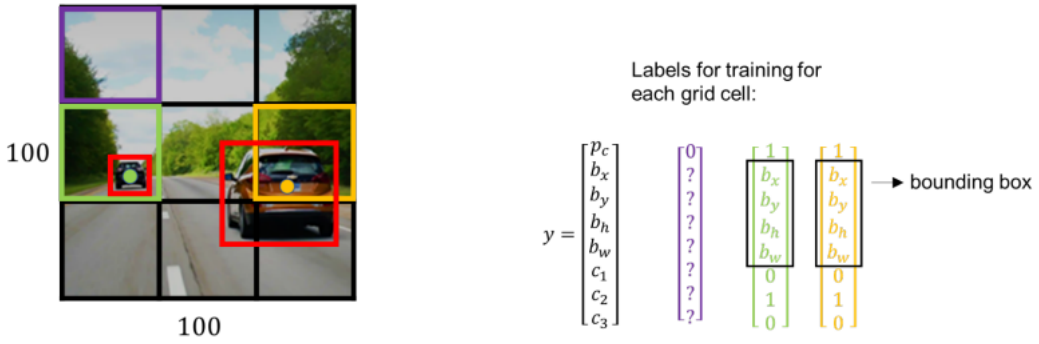


Figura 25: Representación de uso de la técnica

Esta técnica única combina la ventana deslizante con la clasificación y localización, superando las limitaciones de enfoques anteriores.

En YOLO, el objeto se asigna a la celda que contiene el punto central del área de detección, proporcionando una localización más precisa y simplificando el proceso de detección.

Ahora bien, en la arquitectura YOLO, se implementa una estrategia que utiliza dos o más áreas de detección por celda. Aunque en una celda solo se puede detectar un objeto, la introducción de múltiples cajas permite que cada una se "especialice" en aprender diferentes relaciones y tamaños de objetos. Este enfoque polivalente mejora significativamente la capacidad de YOLO para adaptarse a la diversidad de objetos presentes en una imagen. Cada área de detección contribuye con su propia percepción y comprensión, permitiendo que el modelo aborde eficazmente objetos de distintas escalas y formas.

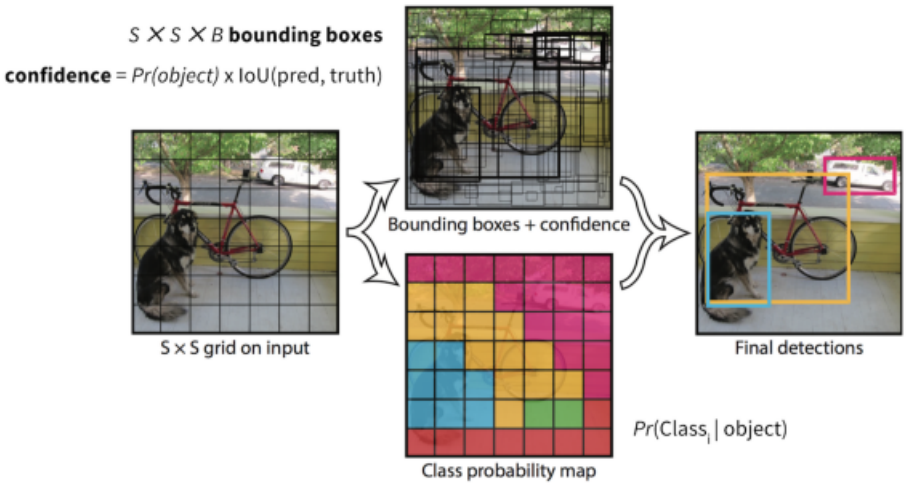


Figura 26: Ejemplo de uso

La dimensión de salida de YOLO se define como $S * S * (5B + K)$, donde S representa el tamaño de la cuadrícula, B denota el número de cajas de detección por celda, y K es la cantidad de clases de objetos a prever. Este diseño compacto y expresivo es fundamental para la eficiencia y versatilidad del algoritmo.

2.15.2. Arquitectura

La arquitectura de YOLO está organizada en tres etapas fundamentales que abarcan desde la extracción de características hasta la generación de predicciones y el posterior post procesamiento:

1. **Backbone:** En la primera etapa, YOLO utiliza una red neuronal preentrenada como su espina dorsal o *backbone*. Ejemplos comunes incluyen DarkNet. Esta fase se encarga de extraer características fun-

damentales de la imagen, aprovechando el conocimiento previo obtenido durante el entrenamiento de la red.

2. **Neck:** La etapa del cuello se caracteriza por capas completamente conectadas que reciben las características extraídas por la espina dorsal. Estas capas se encargan de fusionar y combinar las representaciones de características para capturar información más compleja y contextual. Esta fase actúa como un puente entre la espina dorsal y la capa de salida.

3. **Head:** La última etapa, la cabeza, consiste en la capa de salida que produce las predicciones finales. Aquí se generan las coordenadas de las cajas delimitadoras, las puntuaciones de confianza asociadas a cada detección y las probabilidades de clasificación para las clases de objetos. Además, se aplica el postprocesamiento, que incluye Non-Maximum Suppression (NMS) y el cálculo del Índice de Superposición de Unión (IoU). Estos pasos finales mejoran la precisión y la consistencia de las detecciones, asegurando que se retenga la predicción más confiable y precisa para cada objeto en la imagen.

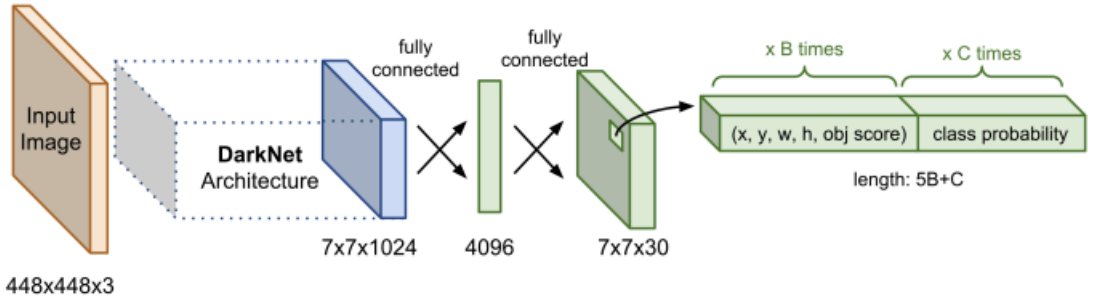


Figura 27: Ejemplo de arquitectura

2.15.3. Entrenamiento - Función Pérdida

La función de pérdida en la arquitectura YOLO se segmenta en dos componentes fundamentales:

1. Pérdida por la predicción del área delimitante (x, y, h, w): L_loc
2. Pérdida por la clasificación del objeto: L_cls

Ambas pérdidas se calculan utilizando el error cuadrático medio (MSE), y cada una de ellas incorpora dos parámetros de escala cruciales: λ_coor y λ_noobj. Estos parámetros influyen en la importancia relativa de cada término de pérdida, permitiendo un ajuste preciso del modelo en función de las metas específicas del entrenamiento.

Pérdida en la localización:

$$L_{loc} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} [(x_i - \hat{x_i})^2 + (y_i - \hat{y_i})^2 + (\sqrt{w_i} - \sqrt{\hat{w_i}})^2 + (\sqrt{h_i} - \sqrt{\hat{h_i}})^2]$$

- **S**: Cantidad de celdas.
- **B**: Cantidad de áreas de delimitación por celda.
- $\mathbf{1_{ij}^{obj}}$:Indica que el área de delimitación j de la celda i es responsable de la predicción.

Pérdida en la clasificación:

$$L_{cls} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} (1_{ij}^{obj} + \lambda_{noobj}(1 - 1_{ij}^{obj}))(C_{ij} - \hat{C_{ij}})^2 + \sum_{i=0}^{S^2} \sum_{c \in C} 1_{ij}^{obj}(p_i(c) - \hat{p_i}(c))^2$$

$$L = L_{loc} + L_{cls}$$

- $\mathbf{1_i^{obj}}$: Indica que existe un objeto en la celda i.
- $\mathbf{1_{ij}^{obj}}$: Indica que el área de delimitación j de la celda i es responsable de la predicción.
- $\mathbf{C_{ij}}$: Confianza de la celda $(Pr(obj) * IoU(pred, truth))$.
- $\hat{\mathbf{C}}_{\mathbf{ij}}$: Confianza de la celda predicha.
- $\mathbf{p_i}(c)$ Probabilidad condicional de una celda i de contener un objeto de clase $c \in C$.
- $\hat{\mathbf{p}}_{\mathbf{i}}(c)$ Probabilidad condicional de una celda i de contener un objeto de clase $c \in C$ predicha por el modelo

El coeficiente de escala de pérdida λ_{noobj} desempeña un papel importante en la prevención de entrenamientos incorrectos, especialmente dado que la mayoría de las celdas en una cuadrícula no contendrán objetos. La determinación del predictor responsable se lleva a cabo mediante el cálculo de la Intersección sobre Unión (IoU). Además, durante el entrenamiento, se emplean cajas de anclaje (anchor boxes) para mejorar y ajustar el rendimiento del predictor.

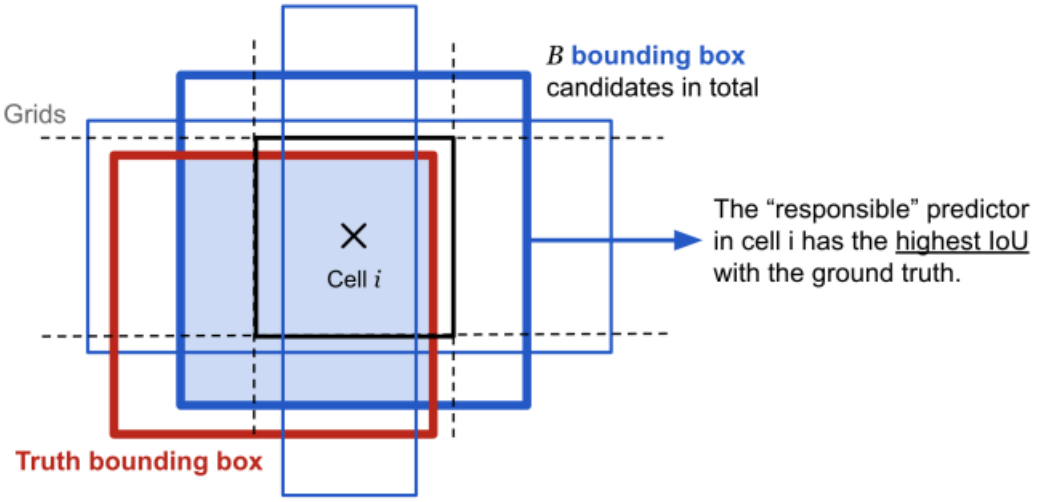


Figura 28: Representación de la técnica

2.15.4. Ventajas

1. **Eficiencia en Tiempo Real:** YOLO es conocido por su eficiencia en tiempo real, ya que realiza la detección de objetos en una sola pasada. Esto lo hace particularmente adecuado para aplicaciones en las que se requiere baja latencia, como en sistemas de vigilancia y conducción autónoma.
2. **Detección de Objetos Múltiples:** YOLO puede detectar múltiples objetos en una imagen, incluso aquellos de diferentes tamaños y clases. Su capacidad para manejar la diversidad en la detección es una de sus fortalezas clave.

3. **Simplicidad y Unificación:** YOLO unifica la clasificación y localización en un solo paso, simplificando el proceso de detección. Su diseño compacto facilita la implementación y el ajuste en diversas aplicaciones.

2.15.5. Desventajas

1. **Precisión en Objetos Pequeños:** YOLO puede tener dificultades al detectar objetos pequeños debido a la resolución limitada de las áreas de detección y la pérdida de detalles en objetos diminutos.
2. **Manejo de Objetos Superpuestos:** En situaciones donde los objetos están superpuestos, YOLO puede enfrentar desafíos para distinguir y delimitar con precisión cada objeto individual.
3. **Problema de Clasificación Desbalanceada:** Si hay una gran variabilidad en la cantidad de instancias de diferentes clases en el conjunto de datos, puede haber desafíos asociados con el equilibrio en la clasificación.
4. **Entrenamiento Sensible a la Inicialización:** El rendimiento de YOLO puede depender en gran medida de una buena inicialización durante el entrenamiento, y un mal ajuste inicial podría afectar la calidad de las predicciones.

Implementación

Resultados

Conclusión

Bibliografía