Cuantificación de incertidumbre en tareas de detección de objetos

Jafed Alejandro Martinez Sanchez

Octubre 2023

1. Motivación

La visión computacional ha sido un área que ha beneficiado enormemente de los avances de aprendizaje profundo. En muchas de las tareas más clásicas (clasificación de imágenes, detección de objetos, reconstrucción 3D, ...), los desempeños han mejorado de forma espectacular. Ahora bien, un punto débil de la gran mayoría de esos sistemas es que carecen de una representación confiable de la incertidumbre asociada a su salida, lo que puede comprometer su integración en sistemas complejos donde la robustez a errores es primordial (por ejemplo, los sistemas de asistencia al manejo automóvil). En este trabajo, para el caso particular de la detección de objetos en imágenes, que es híbrido por naturaleza, mezclando componentes de regresión y de clasificación, se propone evaluar técnicas permitiendo capturar las diferentes formas de incertidumbre asociadas a las estimaciones hechas por aprendizaje máquina, es decir las incertidumbres *aleatoricas* (asociadas a las variaciones "normales" de los datos) y las incertidumbres *epistémicas* (asociadas a la carencia de datos o al uso de un conjunto de entrenamiento específico de un cierto dominio).

2. Metodología

Como modelo determinista de base para nuestras evaluaciones de cuantificación de incertidumbre, se propone utilizar Yolov5, que es una modificación directa al modelo propuesto en [1]. Yolov5 tiene como objetivo la detección de objetos en un conjunto de imágenes con C clases diferentes. Para realizar este proceso, se divide la imagen en una malla de tamaño $S \times S$ de manera que se puedan hacer B predicciones de cuadros delimitadores de objetos por cada una de las mallas en la imagen. Cada uno de estos cuadros delimitadores tiene 5 parámetros (x, y, w, h, p): la posición del centro del cuadro, su ancho, su alto y p, la confianza de la detección. Además, cada elemento del mallado tiene su propio vector de probabilidades para cada una de las C clases. Esta detección se lleva a cabo por medio de una CNN del estilo de U-net, con salidas de tamaño

$S \times S \times B(5+C)$.

Para la evaluación de la incertidumbre *epistémica*, la cual se presenta debido a la carencia de los datos [2], se propone utilizar un modelo de tipo *maskembles*, con M mascaras, como se menciona en [3], en las últimas capas del modelo Yolov5 de manera que para cada imagen se puedan tener al menos M predicciones ($f^m(x)$ con $1 \leq m \leq M$) y formar un ensamble que permite aproximar los resultados de un Ensamble Profundo [4]. De esta manera, la salida de este nuevo modelo llamado **Yolo-mask** es de tamaño $M \times S \times S \times B(5+C)$. Con esta salida, se puede estimar la incertidumbre sobre cada detección utilizando una distribución normal con parámetros:

$$\mu = \frac{1}{M}\sum_{m=1}^{M} f^m(x), \sigma = \frac{1}{M-1}\sum_{m=1}^{M}(f^m(x)-\mu). \tag{1}$$

En el caso de la incertidumbre *aleatoria*, aquella intrínseca en los datos, se propone utilizar el enfoque dado en [2] utilizando el propio modelo para predecir la varianza de cada uno de los cuatro datos espaciales. En otras palabras, a cada una de los B cuadros delimitadores, se asocia su propia varianza como salida del modelo, resultando en una salida de tamaño $M \times S \times S \times B((5+4)+C)$. Con estas nuevas salidas, usamos, para cada variable posicional predicha, una verosimilitud correspondiendo a una distribución normal en la función de costo:

$$L(w) = \sum_i \frac{1}{2}\hat{\sigma}_i^{-2}||y_i - \hat{y}_i||^2 + \frac{1}{2}\log\hat{\sigma}_i^{-2},$$

con $\hat{y}_i, \hat{\sigma}_i$ la predicción realizada por nuestro detector modificado y $\mathbf{y}_i$ el valor *ground truth*.

3. Resultados preliminares

Se realiza la detección de objetos con la base de datos Coco 2017, con los diferentes cambios propuestos en la sección de metodología. En la tabla 1, se muestran los resultados de Average precision (AP), Average Recall (AR) y frames por segundo (FPS) para los modelos de YOLO entrenados: Yolo en formato base; Yolo con *masksembles* y Yolo con *masksembles* y varianza aleatoria. Se puede observar que los modelos modificados tienen un peor rendimiento al modelo base; sin embargo, esto era de esperar ya que las modificaciones al modelo base involucran agregar máscaras a las últimas capas y un parámetro extra a la función de pérdida del modelo, lo cual ralentiza el aprendizaje. Es también un costo para poder recuperar una cuantificación de la incertidumbre. Aun después de estos percances el AP (métrica importante) no se ve muy afectada.

Por otra parte, como se puede apreciar en la figura 1, para las 2 detecciones superiores que corresponden a las salidas de las mascarás 1 y 3 del modelo se puede observar detecciones múltiples o equivocadas para un mismo objeto, como

Modelo	AP	AR	FPS
Yolo base chico	**0.243**	**0.335**	295.51
Yolo mask chico	0.224	0.310	275.61
Yolo mask+var chico	0.213	0.295	305.81
Yolo base mediano	**0.297**	**0.390**	182.71
Yolo mask mediano	0.290	0.382	221.41
Yolo mask+var mediano	0.268	0.346	215.43

Cuadro 1: AP, AR y FPS para los diferentes modelos de YOLO.

en el caso de *calle*, un carro se detectó como *camión* por la mascara 1, y, como *persona* por la mascara 3. La solución para esta problemática es utilizar las salidas de todas las mascaras para seleccionar las detecciones correctas, esto utilizando una modificación del algoritmo de *supresión de no máxima* que permite conservar los elementos necesarios para calcular la incertidumbre epistémica con la ecuación 1. De nuevo, en la figura 1, la tercera fila muestra una versión preliminar de nuestro algoritmo produciendo los resultados de detección al combinar los resultados de las máscaras.

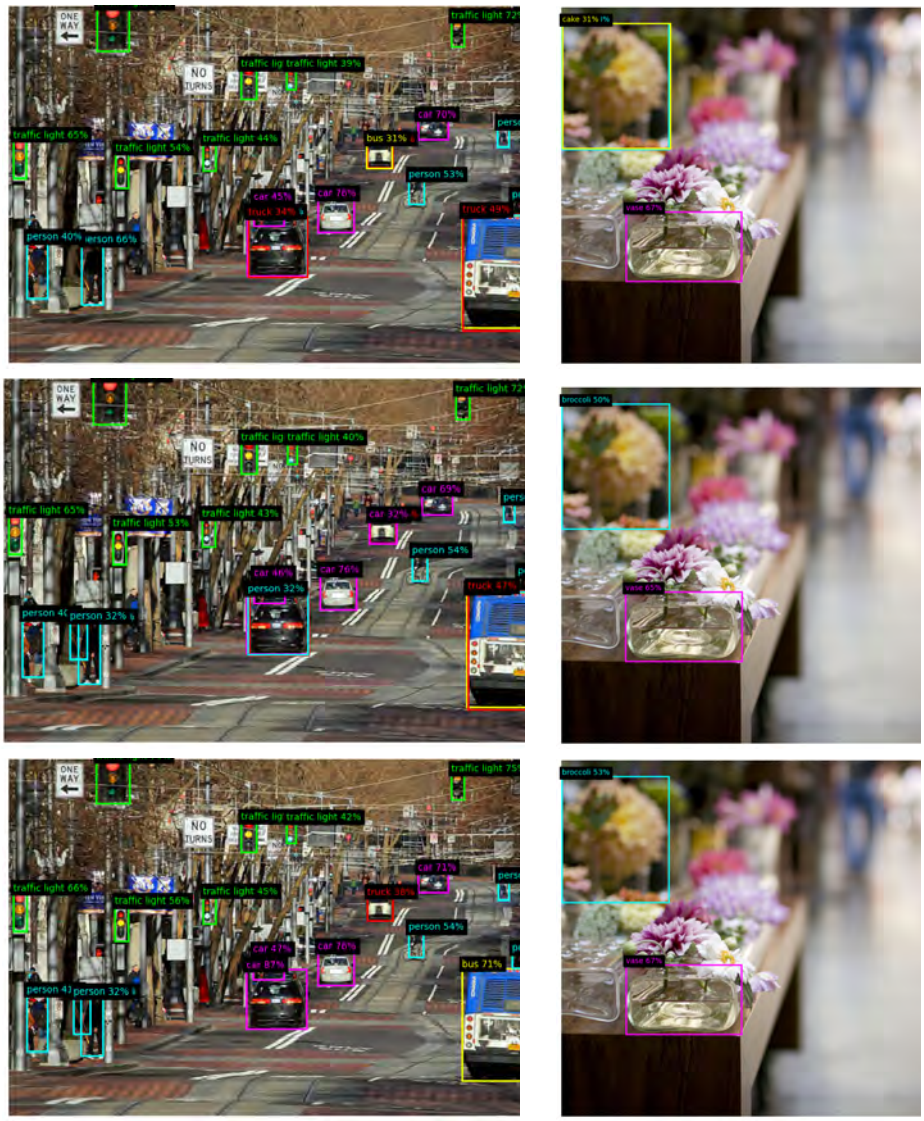
Referencias

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020.

[2] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," 2017.

[3] N. Durasov, T. Bagautdinov, P. Baque, and P. Fua, "Masksembles for uncertainty estimation," 2021.

[4] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.



Calle Flores

Figura 1: De arriba hacia abajo, se muestran las detecciones de la máscara 1, de la máscara 3 y la selección de detecciones por todas las máscaras, para un ejemplo de calle y otro de flores.