RETI NEURALI PER OBJECT DETECTION

Metodi Ancoraggio

Il vantaggio dei metodi basati sull'ancoraggio è che la conoscenza precedente fornita dagli ancoraggi limita le proporzioni e potrebbe impedire forme di rilevamento irregolari. Tuttavia, gli svantaggi (o almeno l'inconveniente) degli ancoraggi possono superare i pregi. In primo luogo, iperparametri come il numero di box di ancoraggio influenzeranno le prestazioni di rilevamento. È meglio evitare questa seccatura. In secondo luogo, poiché tutti gli ancoraggi contribuiscono alla perdita complessiva durante l'addestramento e che solo i relativamente pochi riquadri di delimitazione della verità a terra corrispondono agli ancoraggi positivi, il processo di addestramento potrebbe diventare problematico a causa dei campioni sbilanciati. In terzo luogo, i metodi basati sull'ancoraggio di solito si traducono in rilevamenti sovrapposti e ricorrono a NMS per la post-elaborazione, che può essere errata in alcune situazioni.

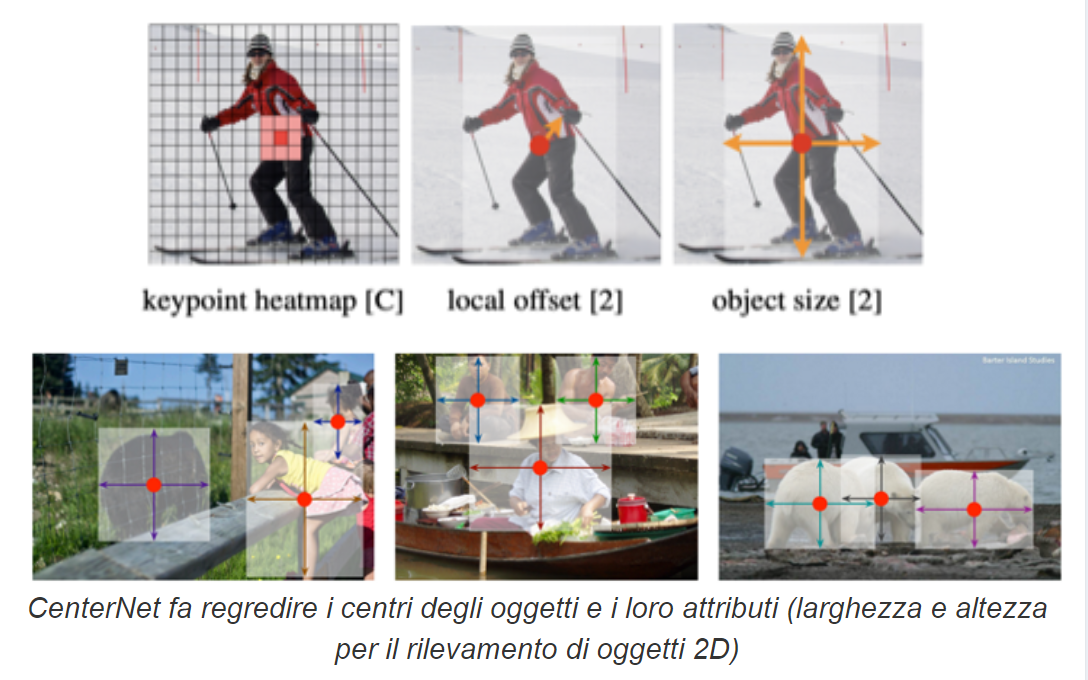
CenterNet

Invece di classificare gli ancoraggi predefiniti in oggetti e regredire le forme del riquadro di delimitazione corrispondente, CenterNet considera gli oggetti come punti e regredisce direttamente i punti centrali degli oggetti e le proprietà corrispondenti, ad es. La dimensione del riquadro di delimitazione, l'offset, la profondità o anche la forma dell'oggetto. Le proprietà dell'oggetto sono altamente personalizzabili a seconda dell'attività e del problema da risolvere.

Vantaggi

* CenterNet è compatibile con i problemi multi-etichetta.
* CenterNet è monostadio e non necessita di ancore o NMS.
* CenterNet fornisce equilibrio tra velocità e precisione.
* CenterNet è altamente personalizzabile ed estensibile.

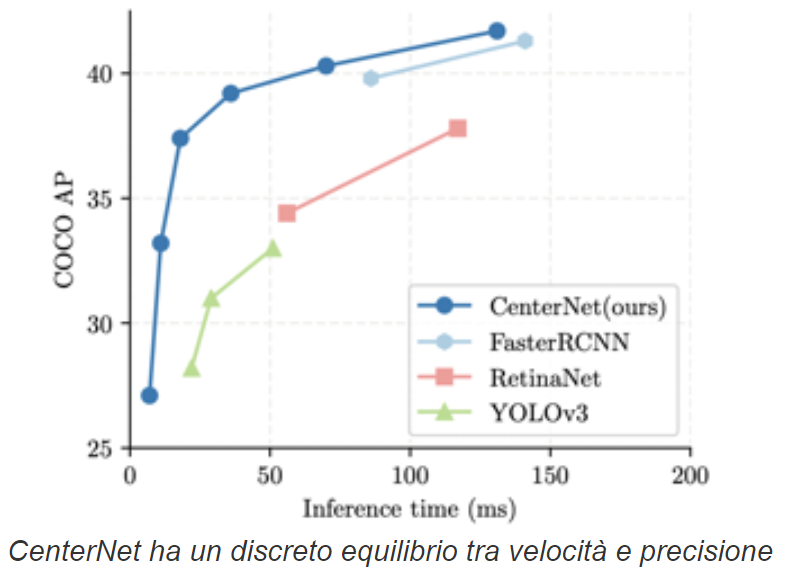
CenterNet trasforma il compito di rilevamento degli oggetti dalla generazione e classificazione delle proposte alla previsione dei centri degli oggetti (punti chiave) e dei loro attributi corrispondenti. Per il rilevamento di oggetti 2D, gli attributi sono la larghezza, l'altezza dell'oggetto, insieme agli offset locali che recuperano la posizione dei pixel nella risoluzione originale dalle mappe di calore sottocampionate. Poiché i punti chiave di generi diversi possono verificarsi nella stessa posizione, CenterNet è naturalmente compatibile con i problemi multi-etichetta.



Con la regressione della mappa termica diretta, gli ancoraggi non sono più necessari. Durante il Training, non è necessario assegnare un oggetto con ancoraggi adeguati. Invece, solo le mappe di calore per l'intera immagine vengono generate come obiettivo di regressione. Quando le mappe di calore dei centri sono previste durante l'inferenza di rete, i picchi locali vengono classificati in base alla risposta o alla confidenza, dove vengono estratti i primi K oggetti. Con le posizioni centrali, gli attributi corrispondenti vengono estratti attraverso i rispettivi canali nella stessa posizione 2D. In questo modo, gli oggetti vengono dedotti e NMS diventa non necessario.

Svantaggi

Rispetto ai rilevatori basati su ancoraggi, CenterNet non si basa sulla conoscenza precedente delle proposte e ciò potrebbe comprometterne il potenziale di prestazioni



Faster-RNN

R-CNN

L’idea che si cela dietro le RCNN è relativamente semplice: iniziano con l’estrarre un insieme di proposte di oggetti (object candidate boxes) usando una selective search. Dunque, per ogni proposta, viene ritagliata un’immagine a dimensione fissa che viene data in pasto ad una CNN addestrata per estrarne le caratteristiche fondamentali. Alla fine vengono usati classificatori lineari SVM per decidere la presenza di un oggetto in ciascuna regione e per riconoscere le categorie degli oggetti trovati.

miglioramento nell’accuratezza del risultato finale, i modelli RCNN presentano una controindicazione importante: il grande numero di features da classificare, risultante dalla sovrapposizione delle tante aree proposte (più di 2’000 aree per ogni immagine), porta il modello a dover elaborare un grande numero di dati, compromettendo le performance dello stesso.

SPP-NET

Nel 2014 K. He et al. hanno proposto Spatial Pyramid Pooling Networks (SPP-Net). Prima di SPP-Net, i modelli di CNN richiedevano un input a dimensione fissa - per esempio immagini di 224x224 pixel per AlexNet. Ciò comportava una perdita d’accuratezza nella classificazione di immagini di dimensione e proporzioni diverse da quelle prefissate, le quali, per poter essere classificate, andavano ridimensionate, con eventuale perdita di dettaglio o deformazione dell’immagine, oppure addirittura tagliate. Il motivo è che, per la loro natura, le CNN sono composte, negli ultimi livelli, da livelli interamente connessi (o fully connected layers), i quali lavorano su di un input di dimensione prefissata.

La grande innovazione di SPP-net sta proprio nell’aver introdotto, tra i livelli convoluzionali e quelli interamente connessi, un livello di Spatial Pyramid Pooling (si veda figura 3.3), il quale permette di mettere insieme le features evidenziate dai livelli convoluzionali e di restituire un output di dimensione prefissata.

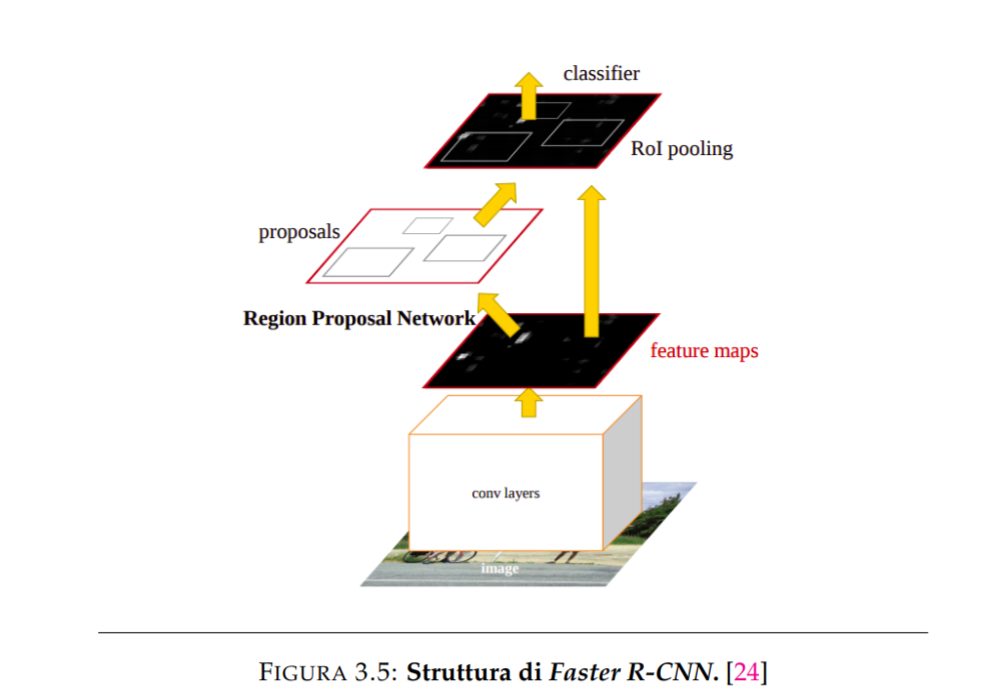
Utilizzando SPP-Net per l’object detection, l’insieme delle aree proposte viene calcolato a partire dalle features estratte dai primi livelli convoluzionali. Successivamente vengono generate delle rappresentazioni di dimensioni prefissata sulle regioni proposte. Queste rappresentazioni possono essere dunque date in pasto ai classificatori

Fast-R-CNN

Fast R-CNN consente di addestrare simultaneamente un riconoscitore ed un disegnatore di bounding box4 all’interno di un unico modello.La loss function di Fast R-CNN infatti tiene conto dell’errore compiuto in ciascuna fase della propagazione

Fast R-CNN prende in input un immagine ed un insieme di object proposal, ossia di aree che si suppone contengano oggetti all’interno dell’immagine; la rete processa l’immagine attraverso una serie di livelli convoluzionali e di max pooling per estrarne le features e produrre una convolutional feature map. A questo punto un livello di pooling, chiamato RoI Pooling Layer (Region of Interest), estrae dalla mappa appena ottenuta un vettore di dimensione prefissata e lo processa attraverso due livelli interamente connessi. Ogni vettore così ottenuto si propaga in due direzioni: in entrambi i casi passa per una serie di livelli interamente connessi; nel primo caso l’output passa poi attraverso un livello softmax per stimare le probabilità, per ciascuna delle K classi di oggetti riconoscibili, che nell’area (vettore) ci sia un oggetto della k-esima classe; nel secondo i livelli producono quattro numeri reali per ciascuna delle K classi di oggetti. Questi valori codificano, per ciascuna classe riconoscibile, il centro e le dimensioni della corrispondente bounding box.

Faster R-CNN



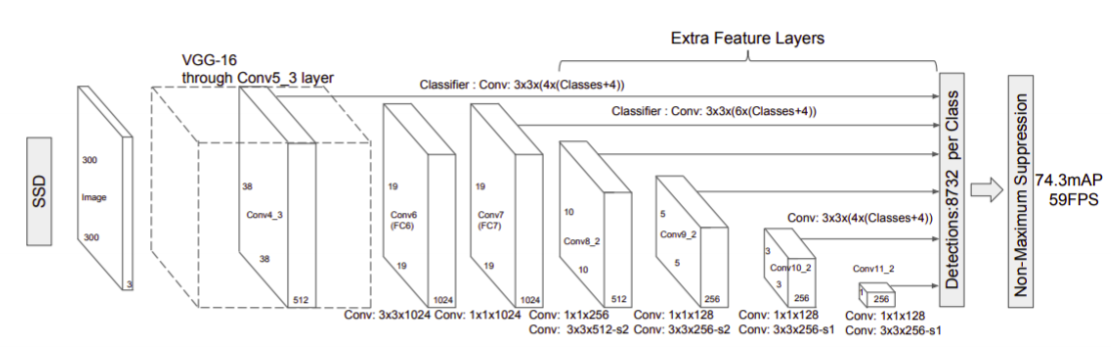
Faster R-CNN è il primo riconoscitore basato su Deep Learning ad essere interamente end-to-end e quasi realtime

La principale innovazione apportata da Faster R-CNN è l’introduzione della rete Region Proposal Network (RPN) che implementa un sistema di region proposal molto efficiente, quasi privo di costi computazionali se paragonato ai modelli antecedenti

L’RPN consiste infatti in una serie di livelli convoluzionali applicati sulle feature map ottenute dai livelli convoluzionali iniziali della rete Fast R-CNN. Per generare le aree proposta, S. Ren et al. optano quindi per integrare una piccola rete alla fine dei livelli convoluzionali di Fast R-CNN. Questa rete prende in input una feature map di dimensione nxn ed è composta da tre livelli convoluzionali. Un primo livello condiviso di dimensione nxn e due livelli "gemelli" interamente connessi di dimensione 1x1 (vedi figura 3.6): fissato l’iperparametro k, il quale denota il numero massimo di proposte da avanzare per ogni locazione, uno dei due livelli gemelli (reg) produce in output le coordinate di k "bounding box"; l’altro livello gemello, (cls) restituisce, per ogni bounding box proposta, le probabilità che in essa ci sia o meno un oggetto.

Sebbene Faster R-CNN superi di molto la velocità di Fast R-CNN, c’è ancora ridondanza nella computazione. Successivamente sono stati infatti introdotti una gran varietà di miglioramenti alla rete.

SSD (Single-Shot Detector)



Il principale contributo di SSD è stato il cambio di prospettiva nei confronti della generazione delle bounding box: a differenza dei modelli precedenti che si preoccupavano di predire con esattezza la locazione di un oggetto all’interno dell’immagine, SSD parte da un insieme di bounding box di default. A partire da questo insieme, SSD predice, per ciascuna di queste bounding box di default, uno scostamento.

Per ciascuna bounding box, traslata dello scostamento predetto, il modello effettua quindi la classificazione. Grazie all’utilizzo di filtri diversi, scelti in base alle proporzioni dell’immagine (i.e. "larghezza"/"altezza"), e grazie all’applicazione simultanea di tali filtri a feature map "multiple", ognuna ottenuta in un punto diverso dei livelli convoluzionali, SSD ottiene anche un’ottima accuratezza nella predizione delle classi di oggetti.

L’evitare la fase di predizione delle bounding box consente al modello SSD di risparmiare molto tempo d’esecuzione. L’utilizzo di feature map diverse invece migliora significativamente l’accuratezza del modello su immagini a bassa risoluzione, specialmente nel riconoscimento di piccoli oggetti.

EfficientNet

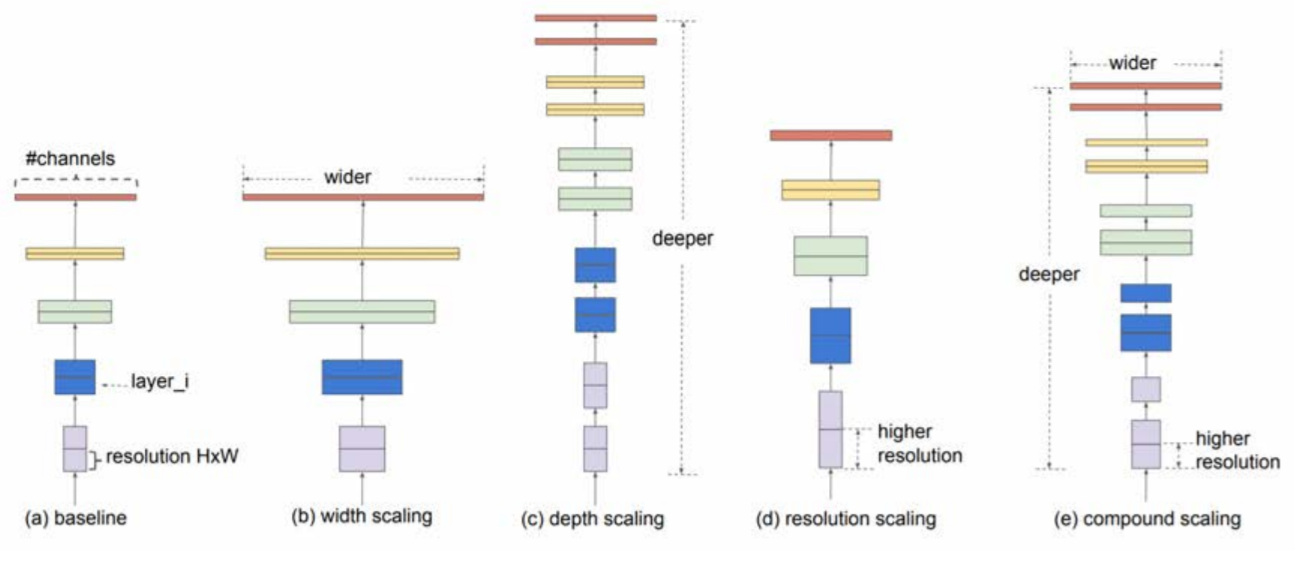
Questa rete non si concentra solo sul miglioramento della precisione, ma anche sull’efficienza dei modelli. Molto spesso il ridimensionamento della rete porta ad un peggioramento delle performance. alcune volte il ridimensionamento, può anche aiutare a migliorare le prestazione di una rete. Come si poteva intuire dai paragrafi precedenti, esistono tre parametri inerenti al ridimensionamento di una CNN:

* **Profondità**: sta ad indicare semplicemente quanto è profonda la rete ed equivale al numero di strati in essa contenuti.
* **Larghezza**: sta a rappresentare quanto è ampia la rete. Una sua misura ad esempio può essere il numero di canali in un livello Conv.
* **Risoluzione**: che indica il dettaglio dell’immagine che viene passata.

Il ridimensionamento del modello, basandosi su una variazione della profondità di questo, è il metodo più comunemente utilizzato per intervenire e risolvere tale problematica. Queste scelte posso derivare da intuizioni che sono diverse per ogni problema affrontato; una rete più profonda ad esempio può acquisire funzionalità (features) più ricche e complesse e generalizzare in maniera sempre più ottimale su nuove attività e campioni del test set. Teoricamente, con più livelli, le prestazioni della rete dovrebbero migliorare, ma praticamente questo non avviene.

il problema del vanishing-gradient è una delle motivazioni più diffuse e comuni che sorgono quando rendiamo la rete troppo profonda. Il ridimensionamento dei pesi (larghezza) di una rete è comunemente usato quando si decide di mantenere il modello il meno esteso possibile. Anche se reti più ampie tendono ad acquisire funzionalità sempre più dettagliate ed efficaci, i modelli più piccoli sono più facili da addestrare in termini di tempo e spazio. Se però tendiamo ad allargare troppo il modello, finiremo per rendere la rete estremamente ampia, con modelli poco profondi, rendendo la rete molto pesante ma con un miglioramento marginale delle prestazioni. Quindi abbiamo capito che non possiamo ne aumentare troppo la larghezza della rete, ne diminuirla in maniera eccessiva. Quindi l’idea che va per la maggiore è quella di collegare e mettere in relazione i ridimensionamenti di larghezza ed ampiezza così da poter cercare un punto di equilibrio.

Ci interessa però anche dare un’occhiata all’ultima tipologia di ridimensionamento ovvero la risoluzione. Possiamo dire che in un’immagine ad alta risoluzione, le caratteristiche sono più dettagliate e quindi a lungo andare queste dovrebbero funzionare meglio nel training a livello di performance ed è uno dei motivi per cui in attività complesse, come il rilevamento degli oggetti, utilizziamo risoluzioni di immagine come 300x300, 512x512 o 600x600. Questo però non è un miglioramento che avviene in maniera lineare; il guadagno di precisione diminuisce molto rapidamente, e ad esempio, l’aumento della risoluzione da 500x500 a 560x560 non comporta incrementi di performance significativi. Ed ecco che, dopo questa breve analisi dei tre punti precedenti, possiamo arrivare alla prima nostra osservazione: il ridimensionamento di qualsiasi dimensione della rete (larghezza, profondità o risoluzione) migliora l’accuratezza, ma il guadagno di precisione diminuisce con modelli sempre più grandi. Sebbene sia possibile manipolare arbitrariamente due o tre dimensioni alla volta, il ridimensionamento arbitrario è piuttosto dispendioso, ed è difficile quindi raggiungere un punto di equilibrio. La maggior parte delle volte, questa variazione manuale determina un’accuratezza ed un’efficienza non ottimali. L’intuizione afferma che all’aumentare della risoluzione delle immagini, anche la profondità e la larghezza della rete dovrebbero essere aumentate. Campi ricettivi più ampi possono acquisire funzioni simili e sempre più complesse che includono più pixel in un’immagine. È fondamentale bilanciare tutte le dimensioni di una rete durante il ridimensionamento delle CNN per ottenere una maggiore precisione ed efficienza; EfficientNet è proprio nata per raggiungere tale obiettivo



mAP (mean Average Prediction)

mAP (mean Average Precision) è una metrica molto diffusa utilizzata  per misurare l’accuratezza dei sistemi di object detection ed indica il valore di precisione medio per il valore di recall compreso tra 0 e 1.

**precision** misura quanto siano accurate le previsioni e si esprime come rapporto tra i True Positive e la somma dei True Positive è dei False positive (risposte ritenute corrette dalla rete ma in realtà errate) . In altri termini andiamo a capire con questa metrica la percentuale di quant volte la rete mi ha restituito falsi positivi.

Precision = TP / (TP + FP)

**Recall** indica quanto il sistema sia in grado di giudicare positivamente e si esprime come rapporto tra i True Positive e la somma dei True Positive e dei False Negative (risposte della rete ritenute negative quando in realtà sono positive).

Recall = TP/ (TP +FN)

**IoU** (Intersection over union) misura la sovrapposizione tra 2 perimetri. Viene usata per misurare quanto il bounding box predetto si sovrappone al ground truth (ovvero il box che contiene l’oggetto reale)

DetectionBoxes\_Precision/mAP calcolata sulle classi con IOU thresholds compreso tra .5 to .95

DetectionBoxes\_Precision/mAP (small) calcolata su piccoli oggetti (area < 32^2 pixels)

DetectionBoxes\_Precision/mAP (medium) calcolata per oggetti di media dimensione [32^2 96^2] pixels)

DetectionBoxes\_Precision/mAP (large) calcolata per oggetti di grande dimensione [96^2 10000^2] pixels)

In COCO mAP, nel calcolo viene utilizzata una definizione AP interpolata a 101 punti. Per COCO, AP è la media su più IoU (il minimo IoU per considerare una corrispondenza positiva). AP@[.5:.95] corrisponde all'AP medio per IoU da 0,5 a 0,95 con una dimensione del passo di 0,05

Lista modelli <https://tfhub.dev/tensorflow/collections/object_detection/1>

Tesi unibo per Faster R-CNN, SSD, Retina-net <https://amslaurea.unibo.it/19637/1/main.pdf>

Tesi EfficientNet <http://tesi.cab.unipd.it/64750/1/pignotti_matteo_tesi.pdf>

Articolo CenterNet

<https://ichi.pro/it/centernet-e-le-sue-varianti-171680071509126>

Articolo EfficientDet

<https://ichi.pro/it/efficientnet-ripensare-la-scalabilita-del-modello-per-le-reti-neurali-convoluzionali-161164828605333>

MaP

<https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173>

<https://www.domsoria.com/2021/04/map-mean-average-precision-e-mar-mean-average-recall-in-object-detection-cosa-sono/>