Analisi di Detectron2

Nel seguente file verranno confrontati alcuni modelli di detectron2 sulla predizioni di soggetti all'interno di immagini.

In particolare verranno confrontati modelli basati su segmentazione panottica e ad istanza.

I modelli faranno uso di backbone differenti come FPN, DC5, C4.

FPN (Feature Pyramid Network)

- Utilizza una struttura piramidale che combina feature estratte a diversi livelli di profondità della backbone.
- È particolarmente utile per compiti come il detection multi-scala, poiché consente di combinare feature a bassa risoluzione (ma ricche di contesto) con feature ad alta risoluzione (che preservano dettagli più fini).
- Le feature vengono propagate verso l'alto attraverso connessioni laterali, creando una gerarchia di feature map con informazioni sia locali che globali.

DC5 (Dilated-C5)

- Questa variante utilizza una dilated convolution (o convoluzione dilatata) sull'ultimo blocco della backbone (C5, il quinto blocco di ResNet).
- L'uso delle convoluzioni dilatate permette di mantenere un'alta risoluzione nelle feature map, senza dover aumentare il numero di parametri o sacrificare il campo visivo della rete.
- Aiuta a preservare più dettagli spaziali rispetto alla versione standard di Faster R-CNN.
- È utile quando si lavora con oggetti di dimensioni variabili ma si vuole evitare il costo computazionale dell'FPN.

C4 (Conv4)

- Il modello C4 utilizza la quarta fase della backbone (C4) per estrarre le feature, senza sfruttare strutture gerarchiche come l'FPN.
- In Faster R-CNN con ResNet, la feature map della fase C4 (ossia il quarto blocco convoluzionale) viene passata direttamente al Region Proposal Network (RPN) e successivamente alla testa di classificazione.
- Questo approccio è più semplice, ma meno efficace nel gestire oggetti di diverse scale rispetto a FPN.

Analisi delle immagini su Validation Set COCO

Le immagini analizzate verranno estrapolate dal validation set di COCO, in modo che ciascun sia già etichettata con i parametri corretti delle istanze da segmentare.

In particolare si vuole capire quanto la maschera predetta si discosti da quella etichettata.

Si effettuerà il test per immagini relative alle super categorie di COCO:

Per ognuna di queste immagini si utilizzeranno backbone FPN, DC5, C4.

Ogni backbone verrà testata con Mask RCNN R50 e R101

FPN verrà testata anche sotto la segmentazione panottica.

R50 ed R101

La differenza principale tra Mask R-CNN con R50 e R101 sta nella backbone utilizzata per l'estrazione delle feature:

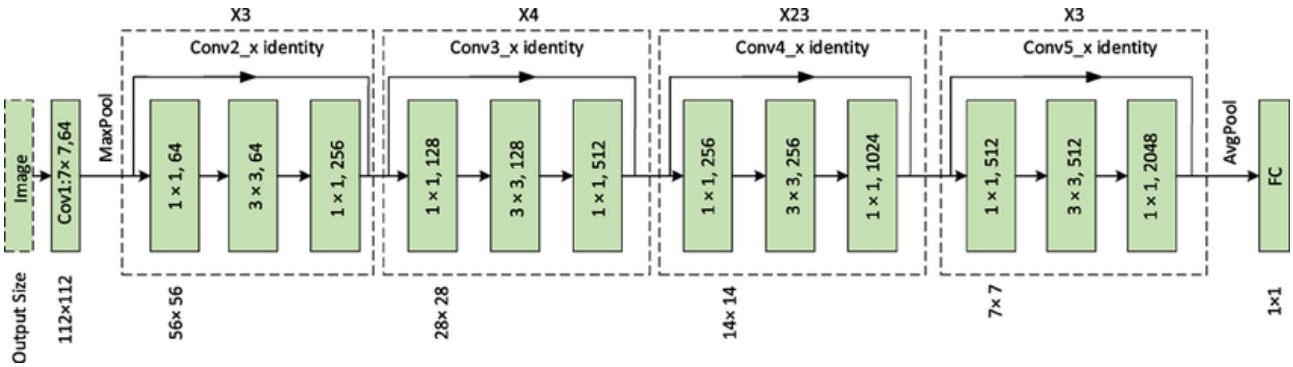
R50 (ResNet-50)

- Ha 50 livelli e utilizza un'architettura più leggera.
- È più veloce e meno costoso in termini di calcolo.
- È adatto per applicazioni in tempo reale o con risorse limitate.

R101 (ResNet-101)

- Ha 101 livelli, con una rete più profonda e capace di catturare feature più complesse.
- Offre una migliore accuratezza, grazie a una capacità di rappresentazione più ricca.
- È più pesante computazionalmente, quindi meno adatto a scenari che richiedono bassa latenza.

(Esempio di tipica struttura ResNet 101)



Ecco un confronto chiaro tra le diverse strutture di Mask R-CNN con ResNet-50 (R50) e ResNet-101 (R101), considerando FPN, DC5 e C4.

Modello	Multi-scala	Risoluzione feature	Velocità	Uso consigliato
Mask R-CNN R50 + FPN	✅ Sì	Medio-alta	🚀 Veloce	Object detection multi-scala, bilanciamento tra precisione e velocità.
Mask R-CNN R101 + FPN	✅ Sì	Medio-alta	🐢 Più lento di R50	Migliore accuratezza rispetto a R50, ma più costoso computazionalmente.
Mask R-CNN R50 + DC5	❌ No	Alta (dilated conv)	🐌 Più lento di FPN	Utile per rilevare piccoli dettagli senza ridurre la risoluzione.
Mask R-CNN R101 + DC5	❌ No	Alta (dilated conv)	🐢 Ancora più lento	Più potente di R50+DC5, ma molto pesante computazionalmente.
Mask R-CNN R50 + C4	❌ No	Media	🚀 Più veloce di DC5	Struttura semplice, meno efficace su oggetti di dimensioni diverse.
Mask R-CNN R101 + C4	❌ No	Media	🐢 Più lento di R50+C4	Migliore accuratezza rispetto a R50+C4, ma meno efficiente.

Risultati dei Test

Per calcolare di quanto la maschera predetta coincide con la GT utilizzeremo la distanza di Housedorff.

Distanza di Housedorff:
Siano X e Y due insiemi non vuoti in uno spazio metrico la distanza di Hausdorff è definita come

$$d_h(X, Y) = max(h(X, Y), h(Y, X))$$

Dove

$$h(X, Y) = sup(inf(d(x, y) : y \in Y) : x \in X)$$

La seguente tabella riporta la distanza di Housedorff in percentuale per ogni istanza trovata in immagine.

Modello	Animal	Elettronics	Food	Person	Road	Road 2
Mask R-CNN R50 + FPN	1 - 12% 2 - 14%	1 - 11%	1 - 9% 2 - 7%	1 - 10% 2 - 4%	1 - 20%	1 - 7%
Mask R-CNN R101 + FPN	1 - 9% 2 - 9%	1 - 11%	1 - 9% 2 - 4%	1 - 10% 2 - 3%	1 - 17%	1 - 7%
Mask R-CNN R50 + DC5	1 - 10% 2 - 14%	1 - 11%	1 - 9% 2 - 7%	1 - 9% 2 - 3%	1 - 15%	1 - 7%
Mask R-CNN R101 + DC5	1 - 9% 2 - 14%	1 - 12%	1 - 9% 2 - 7%	1 - 8% 2 - 4%	1 - 18%	1 - 7%
Mask R-CNN R50 + C4	1 - 11% 2 - 15%	1 - 8%	1 - 9% 2 - 7%	1 - 9% 2 - 4%	1 - 10%	1 - 7%
Mask R-CNN R101 + C4	1 - 10% 2 - 15%	1 - 10%	1 - 9% 2 - 5%	1 - 11% 2 - 5%	1 - 13%	1 - 7%

Media degli errori su ogni modello:

Mask R-CNN R50 + FPN	Mask R-CNN R101 + FPN	Mask R-CNN R50 + DC5	Mask R-CNN R101 + DC5	Mask R-CNN R50 + C4	Mask R-CNN R101 + C4
10,1%	8,7%	9,4%	9,7%	8,8%	9,5%