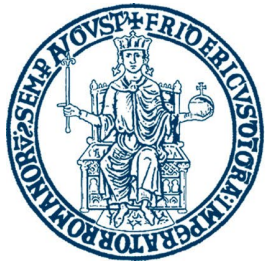


UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

Text Detection tramite Transfer Learning

Christian Perrella
Alessio Giannetti
Carolina Di Donato

Prof.
Luisa Verdoliva
Davide Cozzolino

SCOPO DEL PROGETTO:

ADDESTRAMENTO DI RETINA-NET PER IL RILEVAMENTO DI BOXES DI TESTO

Possibili Applicazioni

Il text detection è alla base di molte applicazioni utili sia per l’utente medio, sia per grandi compagnie.

Quindi il nostro progetto può essere utilizzato come incipit per API più complesse ed efficaci.

IN PARTICOLARE:

- Ricerca visiva. Ad esempio, il recupero e la visualizzazione di immagini che contengono lo stesso testo
- Approfondimenti sui contenuti. Ad esempio, fornire approfondimenti su temi che ricorrono nel testo riconosciuto da fotogrammi. L'applicazione può cercare nel testo riconosciuto contenuti rilevanti, come notizie, punteggi sportivi, numeri di atleti e didascalie
- Spostamenti. Ad esempio, lo sviluppo di un'applicazione in grado di replicare con messaggi audio il contenuto di cartelli stradali o pubblicitari(utile soprattutto per persone ipovedenti)
- Supporto alla sicurezza pubblica e ai trasporti. Ad esempio, il rilevamento dei numeri di targa delle auto dalle immagini delle telecamere di sicurezza
- Filtraggio. Ad esempio, filtrare le informazioni di identificazione personale dalle immagini di documenti

STATE OF THE ART

Google Vision

Microsoft Cognitive Services
Microsoft Azure
Visione Artificiale

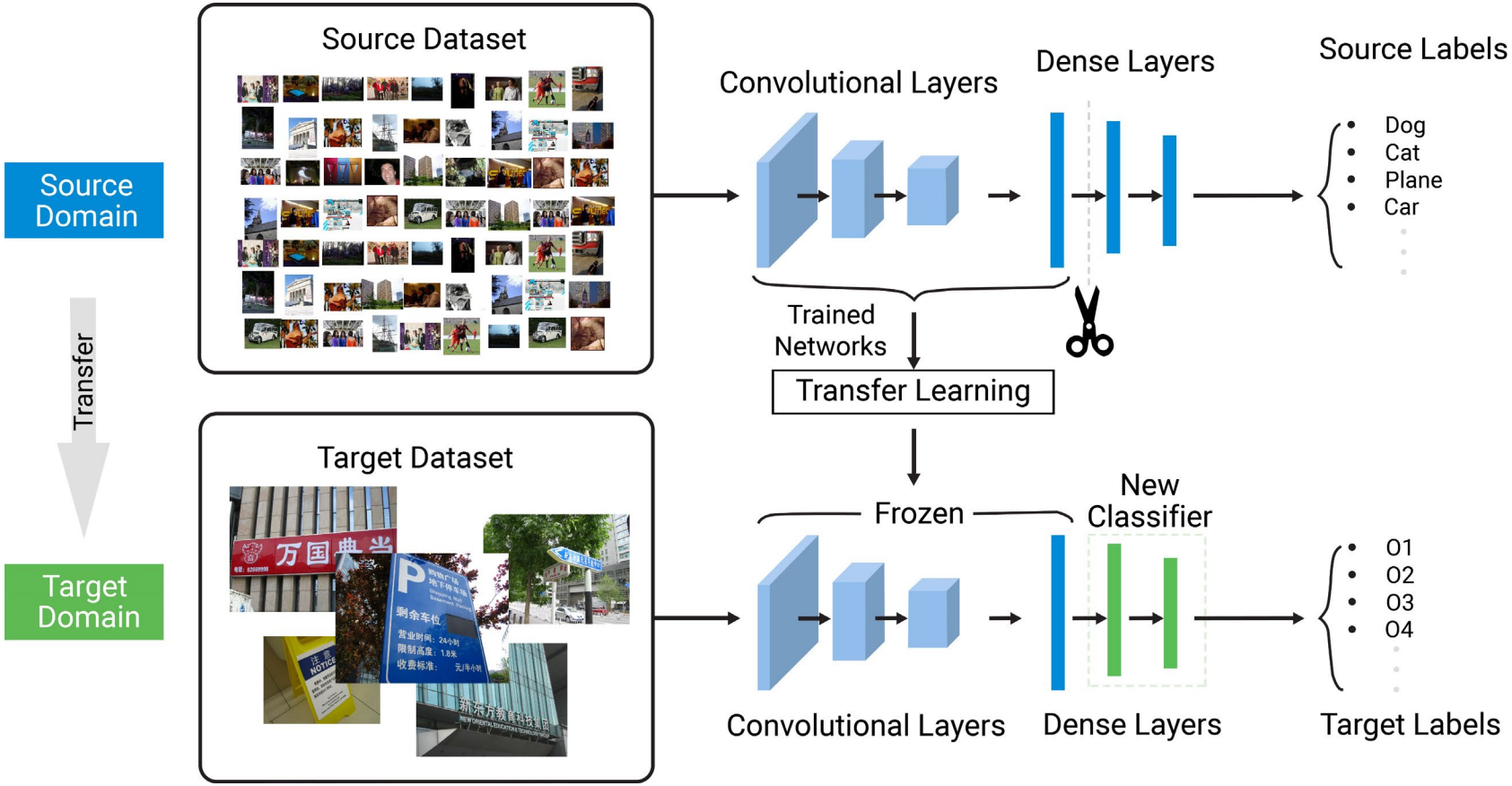
amazon Rekognition

DATASET: MSRA Text Detection 500 Database (MSRA-TD500)

Il **dataset** contiene:

- Immagini scattate in ambienti interni(uffici e centri commerciali) ed esterni (strade);
- Immagini ricorrenti di testo su insegne, targhe, cartelli stradali e cartelloni pubblicitari su sfondi complessi;
- La risoluzione delle immagini varia da 1296x864 a 1920x1280;
- Diverse lingue(inglese, cinese o un mix di entrambe);
- Diversi font, grandezza, colore e orientamento;
- Backgrounds, con vegetazione e pattern ripetuti, che sono difficili da distinguere dal testo.

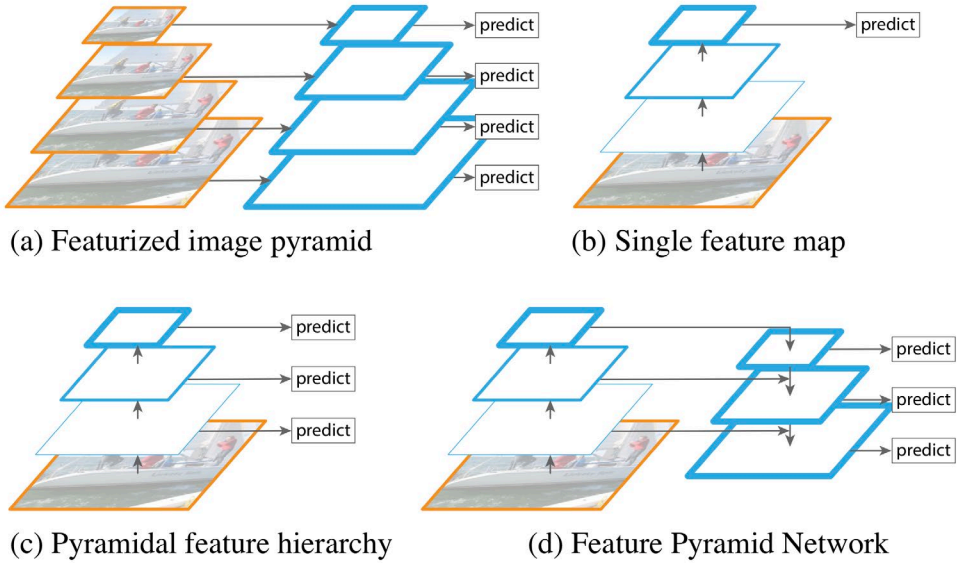
Transfer Learning



Keras RetinaNet

RetinaNet è stata definita sulla base di 2 miglioramenti rispetto agli altri modelli di riconoscimento oggetti a singolo stadio (single stage) – **Feature Pyramid Networks** (FPN) e **Focal Loss**.

(a) mostra la Featurized image pyramid, che è onerosa di risorse perché le features sono ottenute a partire da ogni immagine scalata indipendentemente
(b) Si ottiene una feature map a partire da una singola immagine (scalata), che è possibile usare per fare predizioni veloci
(c) Si usa la gerarchia di feature piramidali a partire dalle feature maps calcolate da una Rete Neurale Convoluzionale
(d) La Feature Pyramid Network (FPN)



Lo scopo è quello di ottenere una Featurized Image pyramid sfruttando la forma piramidale delle feature maps di una rete neurale convoluzionale. Si crea un'architettura in grado di combinare feature a bassa risoluzione e semanticamente valide con feature a alta risoluzione e semanticamente deboli attraverso un percorso top-down e connessioni laterali.

Focal Loss

Focal Loss(FL) è la versione migliorata della Cross-Entropy Loss(CE) ed è stato introdotto per gestire il problema dello squilibrio presente dei modelli di object detection a Singolo stadio (Single Stage object detection models).

I modelli a singolo stadio soffrono del problema dello squilibrio di classi foregroung-background dovuto a eccessivo campionamento di anchor boxes (che potrebbero essere possibili locazioni oggetti).

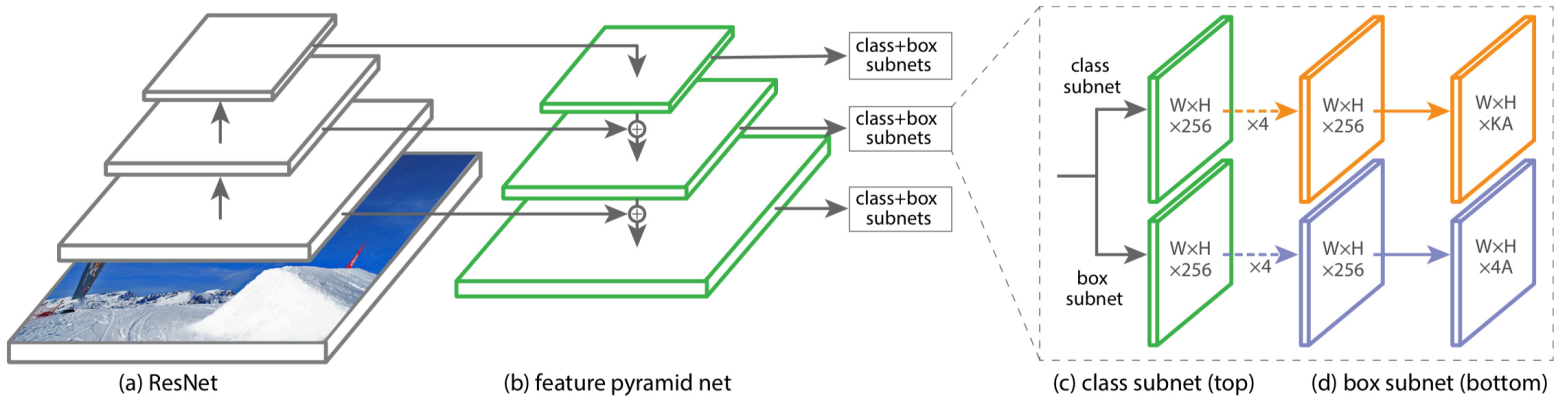
La loss function è una versione della Cross-Entropy Loss scalata di un fattore $(1 - p_t)^\gamma$, che tende a 0 quando aumenta la confidenza della classe corretta. Intuitivamente, il fattore di scala può diminuire il contributo alla loss di campioni semplici durante l'addestramento e far concentrare il modello sui campioni difficili.

$$\mathrm{FL}(p_\mathrm{t}) = -(1 - p_\mathrm{t})^\gamma \log(p_\mathrm{t})$$

Architettura

L’architettura di rete Retinanet usa come backbone l’unione dell’architettura ResNet e di una Feature Pyramid Network per generare una piramide di feature convoluzionale multiscala.
A questo backbone RetinaNet vi aggiunge due sottoreti:

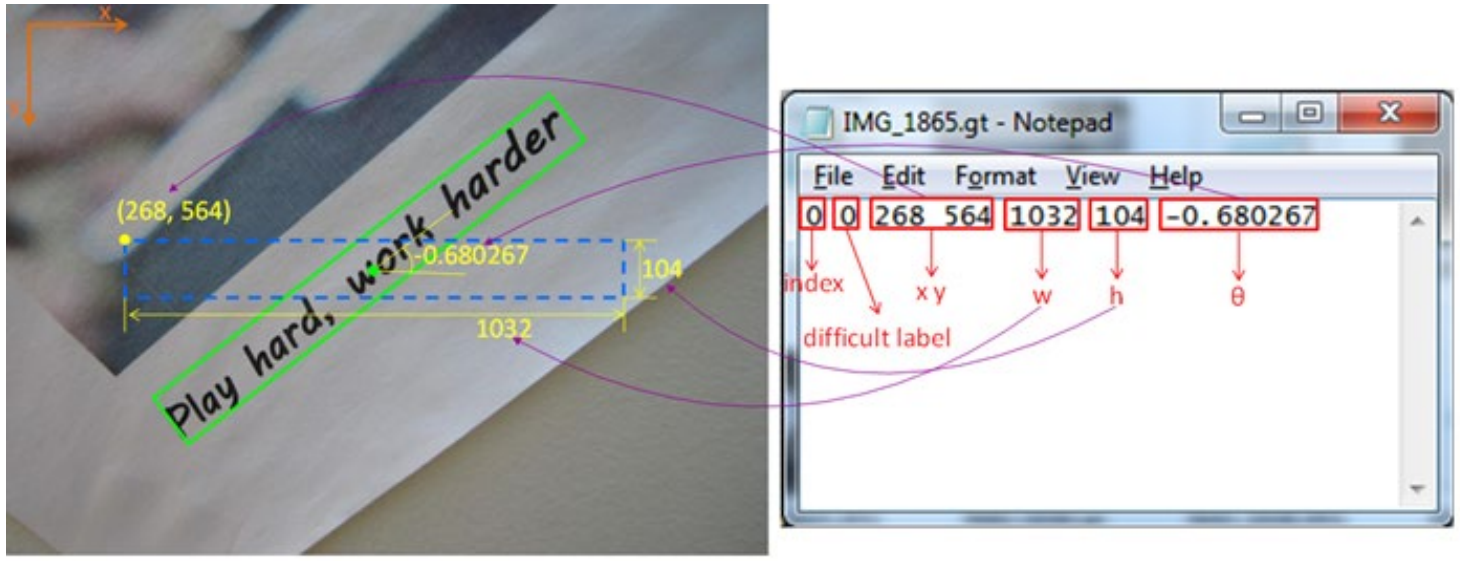
- (c) sottorete per la classificazione di anchor boxes
- (d) sottorete per fare la regressione da anchor boxes a ground-truth object boxes

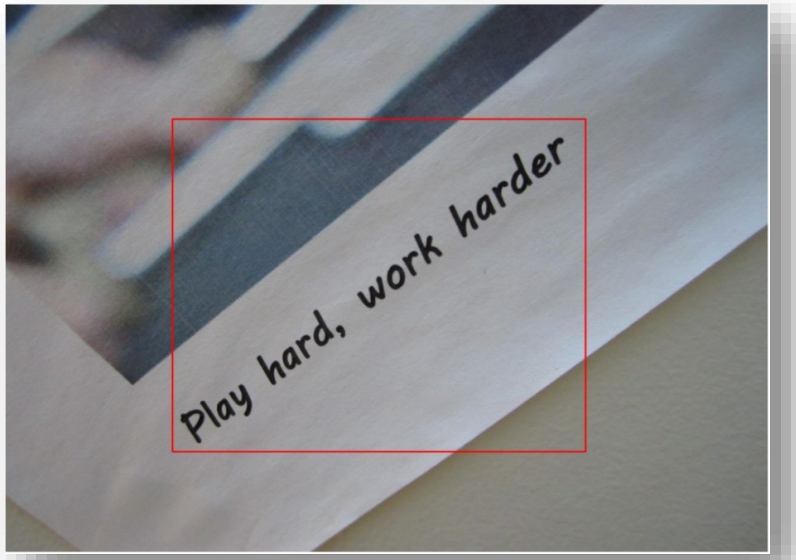


Il nostro approccio al problema

1) Installazione Keras RetinaNet
2) Caricamento del dataset MSRA-TD500
3) Suddivisione del dataset contenente 500 immagini in:
   - Training: 250 immagini
   - Validation: 50 immagini
   - Test: 200 immagini
4) Ottenere CSV dal Dataset
5) Addestrare la rete con opportuni parametri
6) Testing della rete e validazione dei risultati

PREPARAZIONE DEL DATASET





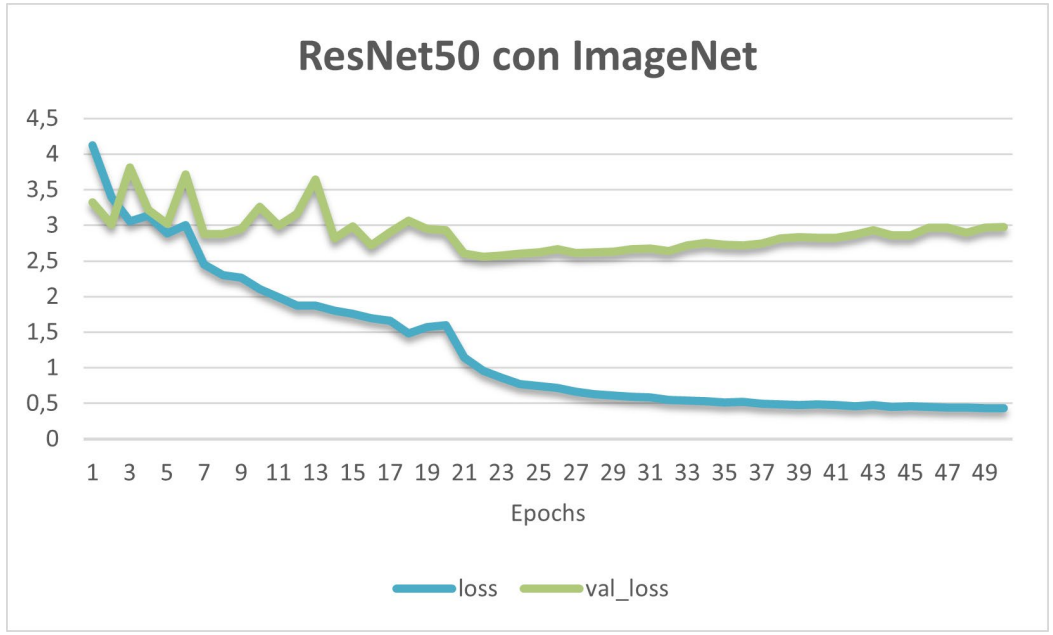
IMG_1865.JPG

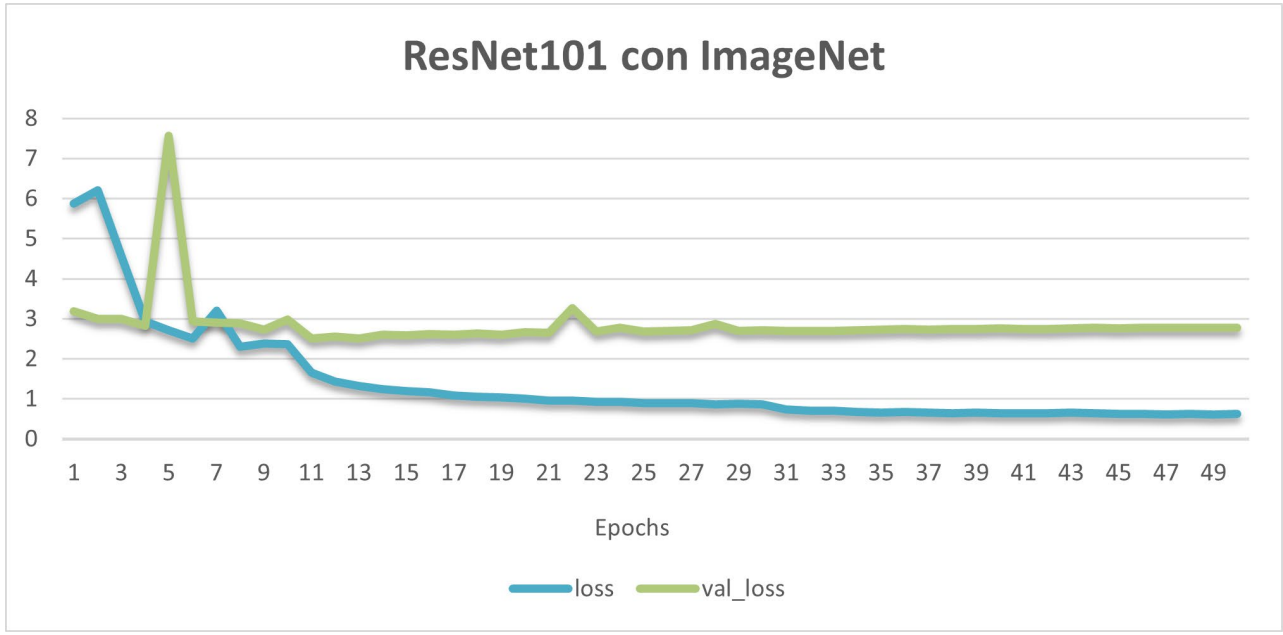
IMG_1865.gt

Path	x1	y1	x2	y2	class
/content/drive/MyDrive/ESM/Progetto/MSRA-TD500/test/IMG_1865.JPG	350	251	1217	980	text
/content/drive/MyDrive/ESM/Progetto/MSRA-TD500/test/IMG_1757.JPG	1000	508	1134	713	text
/content/drive/MyDrive/ESM/Progetto/MSRA-TD500/test/IMG_1757.JPG	650	482	903	565	text
/content/drive/MyDrive/ESM/Progetto/MSRA-TD500/test/IMG_1757.JPG	705	574	798	608	text
/content/drive/MyDrive/ESM/Progetto/MSRA-TD500/test/IMG_1757.JPG	590	613	921	708	text

- **Calcolo dei 4 vertici del box;**
- **Rotazione dei 4 punti;**
- **Applicazione di un algoritmo min/max**

CONFRONTO TRA BACKBONE

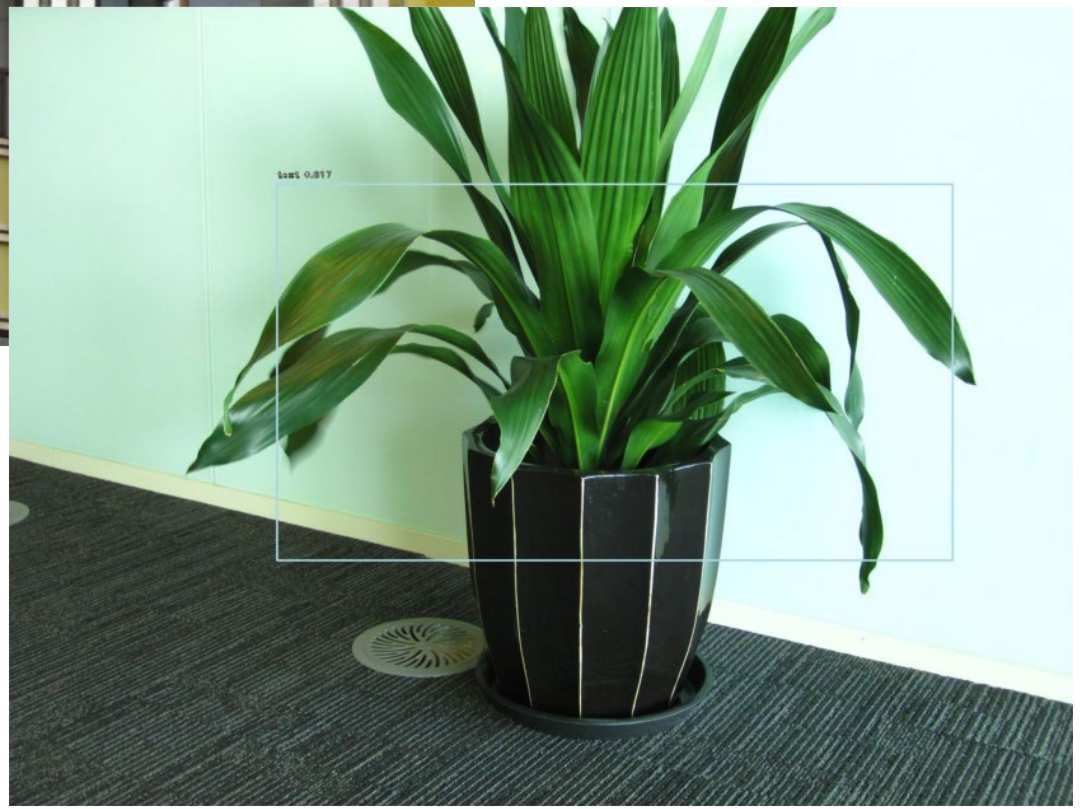




RISULTATI

Epochs	Batch-size	Learning-rate	Steps	Loss
10	1	1E-4	250	3.37
50	1	1E-4	250	2.55
10	1	1E-6	250	3.19
50	1	1E-6	250	2.66
50	1	1E-5	250	1.58

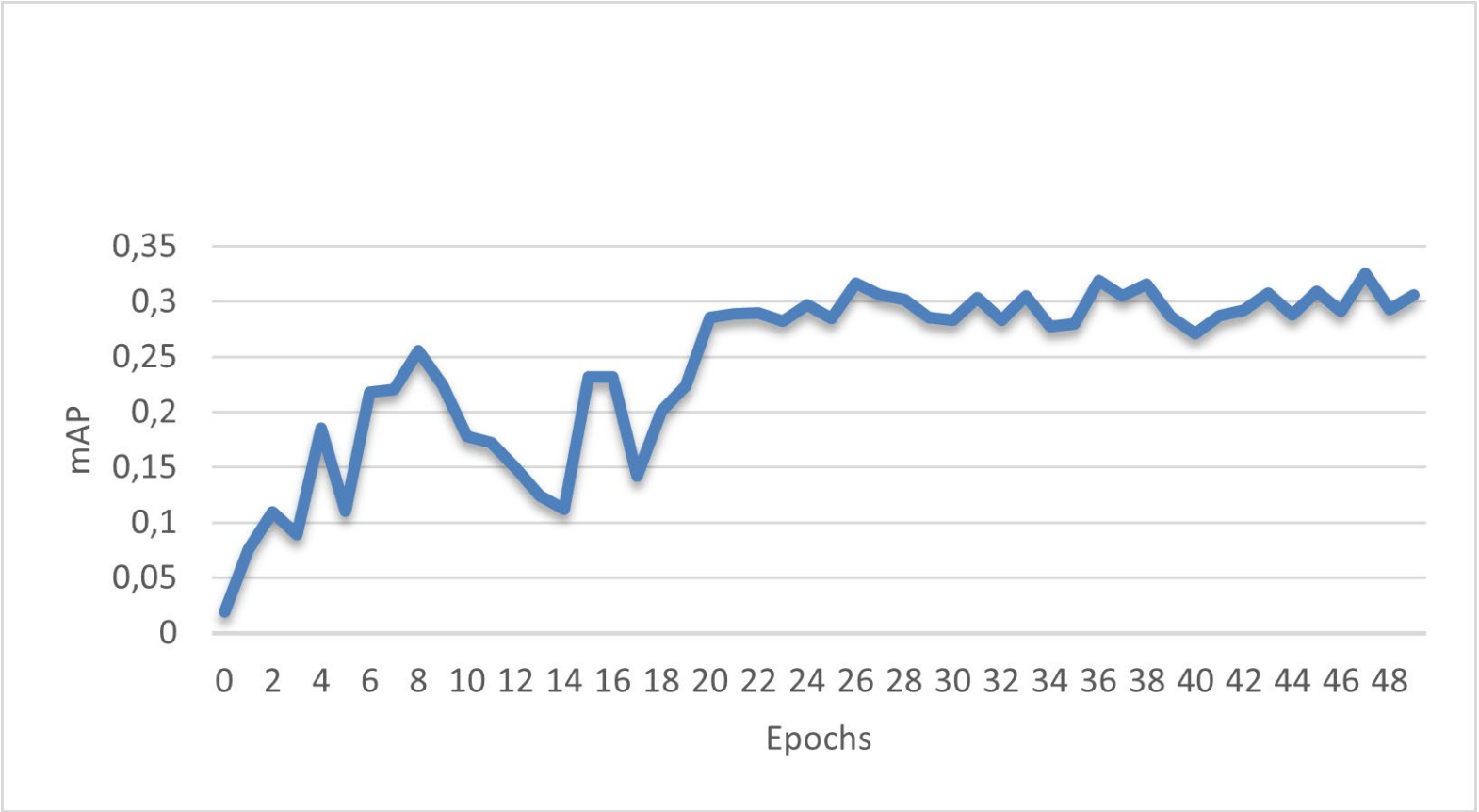
CASI POSITIVI



CASI NEGATIVI

Fase di test

Per la valutazione della nostra rete abbiamo utilizzato il mAP, mean Average Precision. Questo fattore è la media dell'AP, definito come l'area sottesa dalla curva di precisione.



Riferimenti

- RetinaNet Repository: https://github.com/fizyr/keras-retinanet
- C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu., “Detecting Texts of Arbitrary Orientations in Natural Images” Computer Vision and Pattern Recognition (CVPR), 2012:http://www.iapr-tc11.org/mediawiki/index.php/MSRATextDetection500Database(MSRA-TD500)
- T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, “Focal Loss for Dense Object Detection” International Conference on Computer Vision (ICCV), 2017: http://arxiv.org/abs/1708.02002 arXiv:1708.02002
- mAP : https://towardsdatascience.com/map-mean-average-precision-might-confuse-you-5956f1bfa9e2
- Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie . ’’Feature Pyramid Networks for Object Detection.’’: https://arxiv.org/abs/1612.03144