Instance Segmentation for Urban Street Scenes

Francesco Bari
francesco.bari.2@studenti.unipd.it

Eleonora Signor
eleonora.signor@studenti.unipd.it

Abstract

In questo lavoro abbiamo confrontato differenti tecniche di instance segmentation, già esistenti, sul task specifico di Urban Street Scenes. Il nostro interesse verso il topic è nato dal fatto che la segmentazione delle istanze è uno dei compiti fondamentali della visione, tuttavia si presenta ancora complesso e non del tutto esplorato.

1. Introduction

La segmentazione dell'immagine è il processo di separazione di questa in più segmenti, in cui ciascun pixel viene associato a un tipo di oggetto. Esistono due tipologie di segmentazione dell'immagine: la segmentazione semantica e la segmentazione d'istanza. La prima contrassegna oggetti dello stesso tipo con la medesima etichetta di classe; la seconda, d'interesse del nostro lavoro, contrassegna oggetti dello stesso tipo e appartenenti a entità distinte con etichette di classe differenti. L'idea che abbiamo cercato di sviluppare ha riguardato il confronto di diverse metodoligie di instance segmentation. La prima tecnica che abbiamo studiato è stato Mask R-CNN [1], approccio a due stadi. Questa l'abbiamo scelta alla luce del riscontro positivo che ha ricevuto dal mondo della vision research, grazie al suo framework concettualmente semplice e generale, caratterizzato da un rilevamento di oggetti d'immagine efficiente, e dalla generazione in contentemporanea di una maschera di segmentazione di alta qualità per ogni istanza. La tecnica che abbiamo deciso di contraporre a Mask R-CNN è stata BlendMask [2]. Questa è invece una tecnica a uno stadio che si è presentata capace di superare le prestazioni di Mask R-CNN sia a livello di previsione della maschera che per tempo di formazione, su dataset MSCOCO 2017 [3] e LVIS [4]. Ci siamo interessati a verificare se questo rimanesse valido anche su datasets, come *Cityscapes* [5] e *WildDash* [6], appartenenti allo specifico topic di *Urban Street Scenes*, con immagini provenienti dalle strade di tutto il mondo, con molti scenari difficili. Alcuni aspetti che abbiamo testato hanno riguardato cambiamenti di backbone, profondità della rete ResNet [7] e numero di layers congelati. I risultati ottenuti ci hanno confermato quanto già annunciato dai lavori precedenti, generalizzando BlendMask come l'approccio a stadi più promettente. Al termine del nostro lavoro e per non limitare la nostra analisi abbiamo fatto qualche considerazione anche su altre tecniche di instance segmentation quali SOLOv2 [8] e Deep Snake [9].

2. Related Work

2.1. Stage approch

Mask R-CNN [1] è una Rete Neurale Convoluzionale che si presenta all'avanguardia in termini di segmentazione dell'immagine. È la variante di una Rete Neurale Profonda che rileva gli oggetti in un'immagine e vi genera una maschera di segmentazione per ciascuna istanza. Mask R-CNN [1] è l'evoluzione successiva di Faster R-CNN [10], rete neurale convoluzionale basata sulla regione, che produce per ogni oggetto candidato 3 output: l'etichetta di classe, l'offset del riquadro di delimitazione e la maschera dell'oggetto. L'architettura della rete, Figure 1, si compone di una CNN (backbone), che processa l'immagine e estrae la feature map. Dopodichè grazie alla Region Proposal Network vengono presentate le proposte o RoI, sul quale andare a fare riconosciemento del riquadro di delimitazione e previsione delle maschera (head). Inoltre prima di generare l'ouput a ciascuna RoI viene applicato RoIAlign, questo permette di ottenere una maschera dove il layout dell'oggetto viene mantenuto.

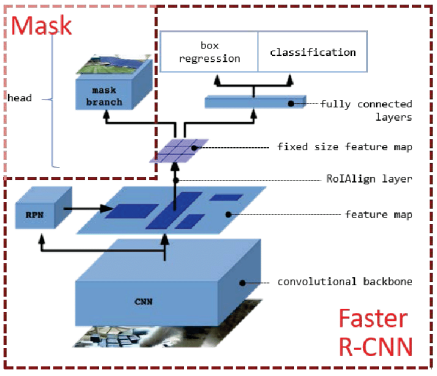

Figure 1. Mask R-CNN architecture. *Source: [11]*

BlendMask [2] deriva dai limiti di Mask R-CNN [1]. Gli autori del paper definiscono come Mask R-CNN [1] vincoli fortemente la velocità e la qualità di generazione delle maschere alle heads, facendo così fatica a trattare scenari complicati e ponendo un limite alla risoluzione delle maschere. Inoltre Mask R-CNN [1] si presenta come un framework poco flessibile per reti multi-task. Hanno così cercato di cobinare strategie di ricerca dall'alto verso il basso e dal basso verso l'alto in FCOS [12], one stage approch, che sembra in grado di superare le controparti a due stadi in termini di precisione. L'architettura di BlendMask [2], Figure 2, si compone da detector network e da una mask branch. Quest'ultima è partizionata in 3 parti: il modulo inferiore che si occupa di prevedere le scores map, chiamate basi; the top layer composto da un singolo strato di convoluzione e da torri, tante quante sono le input features, con il compito di predirre attention instance e un modulo blender che unisce scores con attenzioni.

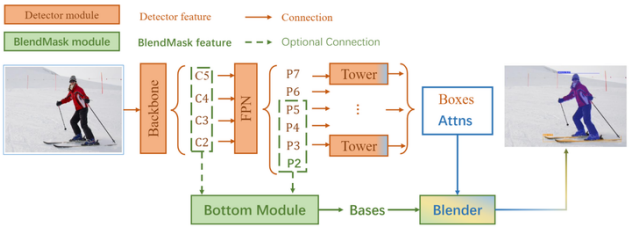

Figure 2. BlendMask architecture. *Source: [2]*

SOLOv2

2.2. Contour-based approach

Deep Snake

3. Dataset

Mostrare qualche immagine contenuta all'interno dei datasets. Al massimo due colonne.
Come sono formati (train, val, test se ci sono), le annotazioni, i json.

3.1. Cityscapes

Ricordarsi che ci sono categorie con frequenza diversa, sarebbe bello mettere un grafico che mostra questa quantificazione

3.2. WildDash

4. Method

Per riuscire a fare un confronto tra le diverse tecniche di instance segmentation, oggetto di questo documento, abbiamo utlizzato i seguenti approcci:

- proceduto con l'implentazione di modelli e successivamente fatto ricorso al metodo sperimentale per la valutazione;
- studiato e analizzato i risultati dei papers.

4.1. Datasets preparation

4.2. Training with fine-tuning

Mask R-CNN La loss utilizzata durante il training è la seguente

$$\min(\mathrm{L}) = \min(\mathrm{L}_{cls} + \mathrm{L}_{box} + \mathrm{L}_{mask})$$

L_{cls} is the classification loss, L_{box} is the bounding-box loss and L_{mask} is the average binary cross-entropy loss.

BlendMask La loss utilizzata durante il training è la seguente

$$\min(\mathrm{L}) = \min(\text{semantic loss})\ [\textbf{?}]$$

.

4.3. Metrics

- AP
- numero di istanze, tempo :: accuracy visiva.confidence-threshold

4.4. Evalutation and inference

5. Experiments and results

In questa sezione descriviamo gli esperimenti che abbiamo eseguito per testare e valutare le tecniche oggetto di questo lavoro. Tali esperimenti gli abbiamo eseguiti al termine delle fasi di studio e codifica.

5.1. Backbone

La seconda serie di esperimenti, che abbiamo compiuto, riguarda la definizione della backbone. Tutte le tecniche a stadi, oggetto del confronto, sono dotate del suddetto modulo inferiore, per cui ci è risultato semplice uniformare le scelte architetturali in modo da poter compiere una valutazione oggettiva. Le tecniche che abbiamo confrontato sono state Mask R-CNN e BlendMask.
Le configurazioni constanti delle reti sono image size ..., numero massimo di iterazioni, learning rate ..., step size a ... e fine-tuning esclusivamente agli ultimi 2 livelli.

Method and architecture	*Cityscapes* AP	*WildDash* AP
Mask R-CNN + ResNet50 + C4 + Base-RCNN-C4		
Mask R-CNN + ResNet50 + DC5 + Base-RCNN-DilatedC5		
Mask R-CNN + ResNet50 + FPN + Base-RCNN-FPN		

Table 1. Backbone Mask R-CNN result.

Per BlendMask, oltre a settare le configurazioni costanti, avvalendoci dei risultati presentati in ... abbiamo settato R

= 56, M = 14, K = 4, sampling method for bottom bases bilinear pooling, interpolation method for top-level attentions bilinear upsampling and semantic loss. Inoltre abbiamo deciso di testare vari tipi di decoder: ProtoNet and DeepLabv3+.

Method and architecture	*Cityscapes* AP	*WildDash* AP
BlendMask with decoder ProtoNet + ResNet50 + FPN + Base-550		
BlendMask with decoder ProtoNet + ResNet50 + deformable convolution + FPN + Base-550		
BlendMask with decoder DeepLabv3+ + ResNet50 + FPN + Base-550		
BlendMask with decoder DeepLabv3+ + ResNet50 + deformable convolution + FPN + Base-550		

Table 2. Backbone BlendMask result.

5.2. Deepness

Una terza serie di esperimenti ha riguardato lo studio della profondità delle reti ResNet.
I parametri di configurazione non definiti in modo esplicito, sono le medesime di quelle riportate nella sezione §5.1.

Method and architecture	*Cityscapes* AP	*WildDash* AP
BlendMask with decoder ProtoNet + ResNet101 + FPN + Base-BlendMask		
BlendMask with decoder ProtoNet + ResNet101 + deformable convolution + FPN + Base-BlendMask		

Table 3. Deepness BlendMask result.

5.3. Freeze levels

Per la quarta serie di esperimenti ci siamo voluti concentrare sul numero di layers da "scongelare" di ResNet durante il re-training dei pesi.
I parametri di configurazione non definiti in modo esplicito, sono le medesime di quelle riportate nella sezione §5.1.

Method and architecture	*Cityscapes* AP	*WildDash* AP
Mask R-CNN + ResNet101 + FPN 1 layers freeze		
Mask R-CNN + ResNet101 + FPN 3 layers freeze		
BlendMask with decoder ProtoNet + ResNet101 + FPN + Base-BlendMask 1 layers freeze		
BlendMask with decoder ProtoNet + ResNet101 + FPN + Base-BlendMask 3 layers freeze		

Table 4. Freeze layers result.

5.4. Own best models

Come ultima serie di esperimenti abbiamo cercato di individuare i modelli migliori, per ciascuna le due tecniche di instance segmentation in esame in questa sezione; tenedo conto della possibilità di allenare ciascun modello solo su una singolo macchina e 1 GPU.
I parametri di configurazione non definiti in modo esplicito, sono le medesime di quelle riportate nella sezione §5.1.

Dataset	method and architecture	AP

Table 5. Own best models result.

6. Conclusion

Al massimo mezza colonna.
BlendMask funziona meglio di Mask RCNN, sia a livello di performance GPU che accuratezza. SOLOv2 sembra avere risultati migliori di Mask R-CNN, ma inferiori a BlendMask. Esistono anche altre tecniche che 'escono' dall'approccio a stadi, per esempio Deep Snake che può presentarsi una valida alternativa a Blender tuttavia da miglirare in futuro. Magari sarebbe possibile un'integrazione tra queste due tecniche.

References

[1] Kaiming He and Georgia Gkioxari and Piotr Dollár and Ross Girshick. Mask R-CNN. CoRR, 2018.

[2] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang and Youliang Yan. BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation. CoRR, 2020.

[3] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. CoRR, 2014.

[4] Agrim Gupta, Piotr Dollár and Ross B. Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. CoRR, 2019.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. CoRR, 2016.

[6] Zendel, Oliver and Honauer, Katrin and Murschitz, Markus and Steininger, Daniel and Dominguez, Gustavo Fernandez. WildDash - Creating Hazard-Aware Benchmarks. Proceedings of the European Conference on Computer Vision, (ECCV), 2018.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition. CoRR, 2015.

[8] Xinlong Wang,Rufeng Zhang,Tao Kong, Lei Li and Chunhua Shen. SOLOv2: Dynamic, Faster and Stronger. CoRR, 2020.

[9] Sida Peng, Wen Jiang, Huaijin Pi, Hujun Bao and Xiaowei Zhou. Deep Snake for Real-Time Instance Segmentation. CoRR, 2020.

[10] Shaoqing Ren, Kaiming He, Ross B. Girshick, Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. CoRR, 2015.

[11] Bienias, Lukasz & n, Juanjo & Nielsen, Line & Alstrøm, Tommy. Insights Into The Behaviour Of Multi-Task Deep Neural Networks For Medical Image Segmentation. 2019.

[12] Zhi Tian, Chunhua Shen, Hao Chen and Tong He. FCOS: Fully Convolutional One-Stage Object Detection. CoRR, 2019.

[13] Xu, Jingyi and Zhang, Zilu and Friedman, Tal and Liang, Yitao and Van den Broeck, Guy. A Semantic Loss Function for Deep Learning with Symbolic Knowledge. Proceedings of the 35th International Conference on Machine Learning, 2018.