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. Esistono differenti approcci di instance segmentation, di seguito ne presentiamo solo una sottoparte, valutandone accuracy e performance su dataset multicategoriali e rivolti a quantificare la robustezza di un algoritmo.

1. Introduction

Mezza colonna

2. Related Work

Mettere le immagini e spiegare le architetture. Al massimo 1 colonna e mezza.
Estendiamo i papers precedenti, cercando di uniformarli, con l'analisi di dataset complessi caratterizzati da molte categorie, differenza di frequenza e immagini complesse che mirano a valutare le tecniche sulla base di illumination challanges.

2.1. Mask R-CNN

Approccio a due stadi, spiegare che deriva da Faster RCNN, usa RoI e RoAlign. Ci sono 3 rami: classificazione, regressione e predizione della maschera

2.2. BlendMask

Approccio a singolo stadio, siegare come funziona il Blender, il modulo inferiore e il livello di attenzione

2.3. SOLOv2

Estensione di Mask R-CNN, fa uso di SGD, del kernel mask G and mask function F

2.4. 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. Stage approach

4.1.1 Preparazione dei dataset

4.1.2 Inferenza

4.1.3 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

$$\min(\mathrm{L}) = \min(\text{semantic loss})\ [4]$$

.

4.2. Contour-based approach

4.3. Metrics

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

4.4. Failure and possibile improvments

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. Stage approach

5.1.1 confidence-threshold

La prima serie di esperimenti che abbiamo svolto hanno riguardato l'inferenza. Abbiamo ritenuto opportuno selezionare tre soglie di confidenza: 0.0, 0.35 e 0.75. Soglie intermedie hanno presentato risultati simili al confidence-threshold di riferimento più vicino.
Abbiamo fatto ricorso a modelli con pesi preadestrati da ImagNet, architettura ResNet 101 e backbone FPN.

Method	value
Mask R-CNN	
SOLOv2	
BlendMask	

Table 1. Inference result. Case none confidence-threshold.

Method	value
Mask R-CNN	
SOLOv2	
BlendMask	

Table 2. Inference result. Case 0.35 confidence-threshold.

Method	value
Mask R-CNN	
SOLOv2	
BlendMask	

Table 3. Inference result. Case 0.75 confidence-threshold.

Inoltre abbiamo notato che in immagini complesse, come a elevata numerosità di oggetti, con differenze di scala o con oggetti deformati nessuna delle tecniche di instance segmentation in analisi, sembra in grado di dare risultati soddisfacenti. TODO: immagini slot di 3. Abbiamo concluso questa prima serie di esperimenti definendo Mask R-CNN, come il modello capace di fornire i migliori risultati di inferenza su modelli standard preaddesstrati.

5.1.2 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 4. 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 5. Backbone BlendMask result.

5.1.3 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 §5.1.2.

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 6. Deepness BlendMask result.

5.2. 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 §5.1.2.

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 7. Freeze layers result.

5.2.1 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 §5.1.2.

Dataset	method and architecture	AP

Table 8. Own best models result.

5.3. Consideration to SOLOv2

SOLOv2 è un metodo a due stadi che ha uso dello stocastic gradient descent per settare i pesi. ... Come si può vedere dal papers di riferimento ottiene risultati migliori rispetto a Mask R-CNN ma che non superano BlandMask (TODO: da approfondire)

5.4. Contour-based approach

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

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

[4] 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.

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