# **12 - BEYOND CLASSIFICATION AND DETECTION**

## Segmentation

A picture containing graphical user interface

Description automatically generated

A close up

Description automatically generated

Nella **semantic segmentation** si tratta di sapere quali pixel appartengono ad una classe. Fa coppia con la instance segmentation dove non si estraggono i bounding box come nella object detection ma tutti i singoli pixel appartenenti ad un oggetto.

*Nota: ci sono algoritmi nati per fare instance segmentation ma che possono essere usati per fare object detection inserendo bounding box.*

Non si ha il concetto di oggetto, ma viene **aggregato tutto alla classe**.

Diagram

Description automatically generated

L’approccio sliding window, chiamato in questo caso Single Pixel Classification, è molto inefficiente.

A picture containing diagram

Description automatically generatedSi sceglie invece un approccio **fully convolutional**, ovvero progettiamo una rete con una serie di livelli convolutivi che vanno a fare la predizione su tutti i pixel in un colpo solo.

* D: numeri di canali
* C: numero di classi che voglio predire
* Mantengo le dimensioni HxW fino alla fine perché voglio mantere tutta l’immagine
* Nella rete faccio variare solo la profondità che alla fine dovrà essere pari al nr di classi

Vantaggi:

* Rete totalmente convolutiva
  + Tutto in un colpo solo
  + Filtri -> alcune cose imparate per una determinata regione di un’immagine serve anche in un’altra regione

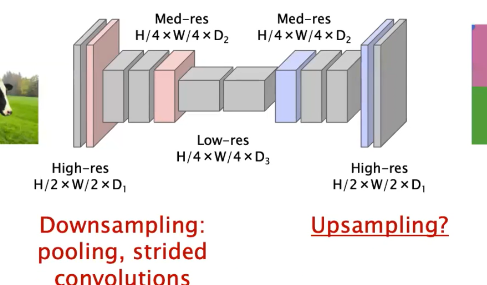
Ha dei problemi:

* Uso la stessa dimensione dell’input HxW nelle convoluzioni 🡪 oneroso per il calcolo e non scala con l’input size

**La soluzione adottata è il modello encoder-decoder**

Diagram

Description automatically generated

All’interno della rete si compatta la rappresentazione usando pooling/ strided convolution (filtro a passi più grandi) (l\*h), nr di filtri diversi (profondità).

In pratica abbiamo una rete con dei livelli convolutivi chiamati di downsampling ottenuti con il pooling o con le strided convolutions, seguiti da dei livelli di upsampling per ri-incrementare la classificazione per fornire l’output.

**In-network upsampling: Unpooling**

Voglio far diventare la matrice 2\*2 -> 4\*4.

Diagram

Description automatically generated

* + - Le celle adiacenti prendono quel valore.
    - Riporto una sola volta il valore, il resto a 0

**Si tiene traccia delle localizzazioni spaziali** di dove si trovano i valori per collegare la fase di downsampling con quella di upsampling.

È importante notare che non ci sono parametri.

Varie tipologie:

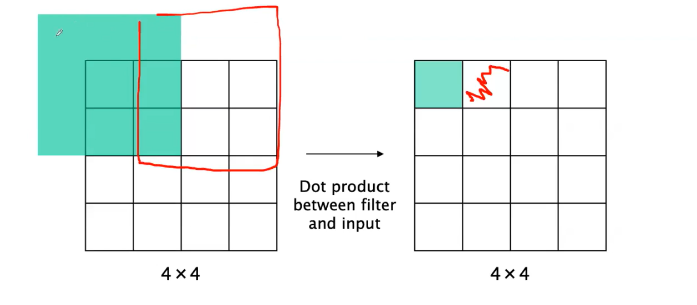
### Max unpooling

Diagram

Description automatically generated

Alto sx: devo usare 1 per popolare la regione, lo inserisco nella posizione in cui era stato trovato il massimo -> il massimo di quel quadrato deve essere quel quadratino.

### Learnable upsampling

Chart, bar chart, box and whisker chart

Description automatically generated

Stride convolution in cui compatto l’informazione ma imparo anche qualcosa.

Chart, bar chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

In questo modo faccio un downsampling

### Transpose convolution

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Immagine che contiene diagramma, linea, Rettangolo, quadrato

Descrizione generata automaticamenteDiagram

Description automatically generated

Dove c’è sovrapposizione, sommo

## **SegNet (Reti Encoder-Decoder o AutoEncoder)**

Posso usare vgg-16, resnet, etc.

Solitamente vengono usati i primi 13 livelli di VGG-16 buttando i FC (così restano solo i convolutivi), riducendo il numero di parametri da 100 milioni a 10 milioni. Alla fine c’è un softmax multiclasse che produce l’immagine segmentata.

Non usa informazioni temporali -> lavora immagine per immagine, dunque ha una latenza bassa.

A picture containing stationary, implement, pencil

Description automatically generated

(linee orrizzontali, collegano le operazioni di pooling della fase di downsampling alle operazioni di unpooling della fase di upsampling)

Nota: non necessariamente simmetrica.

**ENet (EfficientNet)**

Versione migliore di SegNet

Table

Description automatically generated

**U-Net**

Chart, box and whisker chart

Description automatically generated

Si comporta molto bene nell’identificare e separare oggetti che vengono a contatto.

Portano avanti nella rete l’informazione attraverso dei modelli di copy and crop -> copie dell’immagine di partenza tagliata (con la speranza che questo aiuti la rete a convergere)

La complessità in questi modelli è produrre un insieme di immagini e più o meno manualmente tracciare i contorni delle varie istanze -> istance segmentation.

**Instance segmentation**

A picture containing grass, outdoor, person, photo

Description automatically generatedprodurre queste annotazioni è molto complicato (separare ciclista dalla bici)

## **Mask R-CNN**

Diagram

Description automatically generatedUsa il meglio della detection usando Faster R-CNN e usa il Fully Convolutional per la segmentazione.

* tira fuori delle ROI su cui fa classificazione degli oggetti
* tira fuori i bounding box degli oggetti
* usa un’altra parte per tirare fuori la mask *(fully convolutional segmentation)* 🡪 **identificare gli oggetti.**

Questo viene fatto per tutti gli oggetti.

*Nota: maschera= insieme dei pixel che rappresentano l’oggetto.*

*Rispetto alle Faster R-CNN c’è una parte di ROI pooling che prende le ROI estratte e va a proiettarle in regioni di dimensione fissa e per farlo va a fare una media dei valori all’interno delle varie feature maps.*

**ROI Align vs Pooling**

Calendar

Description automatically generated

pooling-> per passare da una dimensione all’altra

Align: filtraggio bi-dimensionale che dovrebbe permettere di mantenere l’informazione più inalterata

Viene applicata una bilinear interpolation

*Risultati:*

A picture containing outdoor, grass, person, photo

Description automatically generated

Circa 5 frame al secondo ovvero 200 ms per immagine

Questa rete può essere usata anche per estrapolare i landmark

A picture containing indoor, toy, child, photo

Description automatically generated

## FACE VERIFICATION AND RECOGNITION

* Graphical user interface, text, application, email

  Description automatically generateddata un’immagine di input, dire se quell’immagine corrisponde ad una determinata persona
* Dato un data-set che contiene n persone, identificare di che persona si tratti

**One shot learning**

Si chiamano one-shot learning perché si ha un’unica immagine da cui imparare il riconoscimento per il seguito. Questo perché, se dovessi voler aggiungere un nuovo soggetto con gli approcci standard dovrei riallenare di nuovo tutta la rete, ri-addestrare l’algoritmo.

#### **Learning a “similarity function”**

Mi dice se due immagini sono simili o no, in base ad una determinata soglia.

Date due immagini si definisce d(img1, img2) come il grado di differenza tra queste due e impostando una soglia posso dire se sono la stessa o meno.

Immagine che contiene Viso umano, testo, schermata, persona

Descrizione generata automaticamente

#### **Siamese network**

Prendiamo una rete e buttiamo l’ultima parte di classificazione

A picture containing furniture, table, game

Description automatically generated

Il nuovo ultimo livello sarebbe l’encoding x(1) e contiene le feature calcolate dai livelli precedenti e lo chiamiamo f(x(1)).

Prendiamo un’altra immagine e prendiamo la stessa rete e abbiamo f(x(2)).

A picture containing clock

Description automatically generated

Se possiamo assumere che gli encoding sono una buona rappresentazione delle immagini allora si può definire d come sopra.

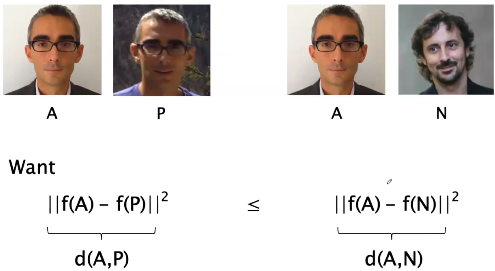
**Goal of learning**

Shape

Description automatically generated

#### **Triplet loss**

Nella realtà abbiamo a che fare con una rete siamese a cui passiamo delle immagini e dobbiamo confrontare essenzialmente l’anchor con un esempio positivo e con uno negativo.



La nostra rete potrebbe imparare un encoding per cui a tutte le immagini associo un vettore di zero per avere la regola sempre verificata.

Quindi **si introduce un margine** che è un hyperparametro della rete

Immagine che contiene Viso umano, occhiali, persona, uomo

Descrizione generata automaticamente

Definisco questo nuovo loss per la tripletta

Text

Description automatically generated

Per ogni soggetto serve un esempio positivo e uno negativo e bisogna fornire esempi difficili per addestrare bene la rete.

Immagine che contiene Viso umano, sorriso, occhiali, collage

Descrizione generata automaticamente

Servono circa 10000 immagini di 1000 persone però una volta addestrata si può usare per problemi di one shot learning.

Graphical user interface, text, application, chat or text message

Description automatically generatedShape

Description automatically generated**Learning similiarity with binary classificationceh**

1 se rappresenta la stessa persona, 0 altrimenti

## **Neural style transfer**

Graphical user interface, application

Description automatically generated

* Immagine che contiene le informazioni della prima immagine con lo stile della seconda

**Visualize what network is learning**

A picture containing text

Description automatically generated

Presa una rete convolutiva, cosa impara la rete?

* Prendiamo un neurone al primo livello, quali sono le immagini che massimizzano l’attivazione di que neurone?

Se andiamo a prendere per nove unità del layer 1

A picture containing application

Description automatically generated

Ognuna di esse viene attivata da un certo tipo di elemento di basso livello presente nell’immagine.

A picture containing photo, many, different, bunch

Description automatically generated

#### **Neural style transfer cost function**

La rete dovrebbe minimizzare la funzione di costo tra l’immagine di contenuto e quella da realizzare e al tempo stesso, cercare di minimizzare la funzione di costo tra immagine di stile e l’immagine da generare.

Graphical user interface, application, website

Description automatically generated

Alfa e beta sono hyperparametri.

**Generate the output image**

A picture containing text, newspaper

Description automatically generated

* Man mano che la rete impara, prova a realizzare un’immagine che assomigli ad entrambe le immagini.

Si inizia settando G randomica e si definisce una funzione di costo settando i parametri per allenare la rete.

#### **Content cost function component**

Considerando il costo della funzione di content utilizzo un layer intermedio per valutarlo in quanto alla fine della rete otterei semplicemente un output quasi identico al content.

Si prende una rete convolutiva pre-addestrata (es. VGG), e si immagini di avere le attivazioni calcolate per quel livello calcolate sul content e sull’immagine generata.

**Se le attivazioni sono simili allora le immagini sono simili** e misuro il costo come differenza tra queste.

A picture containing indoor, bird

Description automatically generated

### **Misurare lo stile di un’immagine**

A picture containing shape

Description automatically generatedCome prima si prende un livello intermedio per misurare lo stile di un’immagine che viene definito come la correlazione tra le attivazioni sui diversi canali.

Se c’è correlazione positiva tra i canali allora le immagini hanno lo stesso stile.

A picture containing diagram

Description automatically generated

#### Style matrix

Si crea la matrice di stile che mi dice quanto sono correlate in profondità le attivazioni di ogni canale.

k è l’indice di profondità che si muove lungo i canali.

L’elemento della matrice è dato dalla somma dei prodotti delle attivazioni. Questo mi dice quanto sono correlate in profondità le attivazioni.

Text

Description automatically generated

Concettualmente è una covarianza non normalizzata e la matrice che contiene tali elementi è una matrice di GRAM.

Lo si fa sia per lo stile che per l’immagine generata

Diagram, schematic

Description automatically generated

##### Style cost function component

Si definisce quindi la funzione di costo dello stile come la differenza tra le due matrici di stile. In realtà si prende la norma di Flobenius dove si estraggono le informazioni da più livelli.

Graphical user interface, text, application

Description automatically generated

Dando un peso ad ogni livello come parametro.

Lezione 15/11