An effectiveness method for fashion parsing task

Student: Khuc Giang Sinh
Supervisor: Do Thai Giang

Table of content

INTRODUCTION

Introduction

E-commercial site try to come up with new features technology to increase customer interaction.

For fashion majors, many application have been known: Searching product, recommended, visual try-on,...

E-commercial applications for fashion products

In-Shop Clothes Retrieval

query image

retrieved

Fashion Compatibility and Recommendation

Diverse Fashion Collocations

Given 1 query item, generate fashion sets of diverse styles and flexible length

Dataset: Polyvore

Fashion Virtual Try-on

query item

Analogous

FASHION PARSING

To deploy that applications, we need to solve fundamental task: clothing classifying, localization, landmark detection,..

It must first recognize the human body component of the input image in order to determine where the clothing area is located and then synthesise clothes in that location.

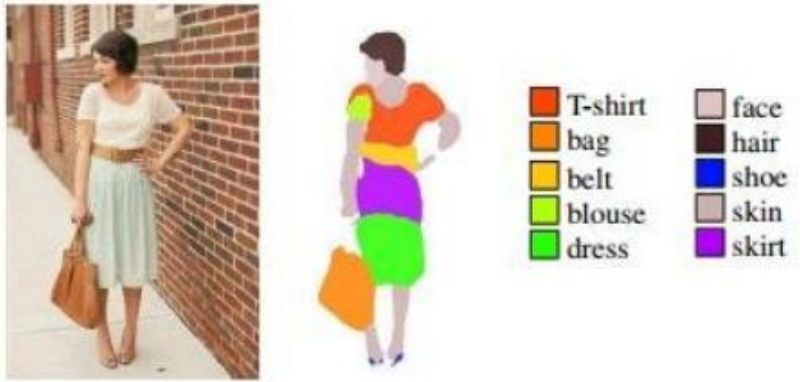


Figure 1: Example of Fashion parsing task

THE PROBLEM

There are many clothes/products in one photo

Large intra-class variance: external conditions such as lighting, background noise, clothing shape, distortions and deviations between user domain and store domain images (shop domain) makes images from domains relatively/absolutely difficult to parsing.

Minor inter-class variance: The factors that make an image distinguishable from other classes are quite small (the long skirt image may be mistaken for a slightly shorter skirt,...)

RELATED WORKS

Yamaguchi were the first to work on fashion parsing. By mutually improving two difficulties, they utilised the link between garment parsing and human posture estimation.

In 2019, human parsing problems were considered using hierarchical graphs

DeepFashion2 challenge 2020 with top method: Aggreation and Finetuning, DeepMark,...

CONTRIBUTIONS

We apply a channel attention module on the backbone of the mask-rcnn model and test on DeepFashion2 dataset to know how this method affects its result.

We show the failure case of mask-rcnn model

METHODOLOGY

GENERAL ARCHITECTURE

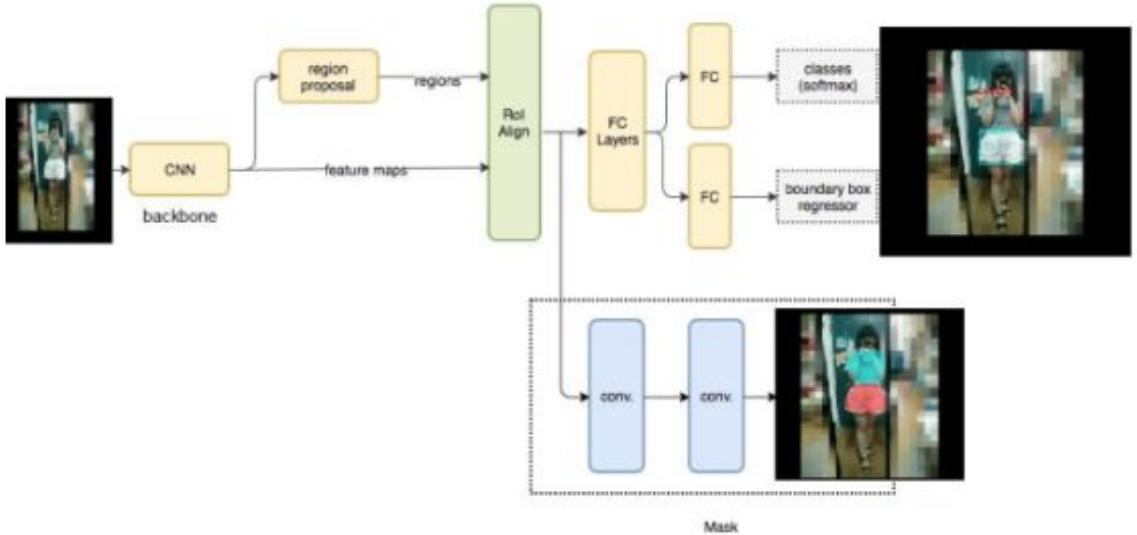


Figure 3.1: General our Mask-RCNN architecture

BACKBONE FEATURE EXTRACTION

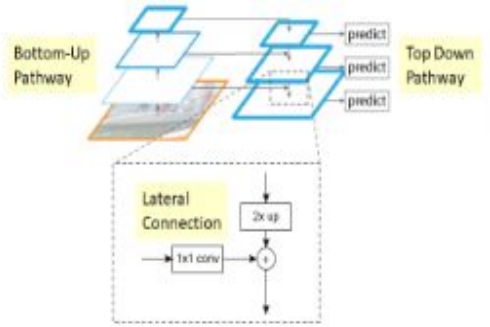


Figure 3.2: FPN architecture

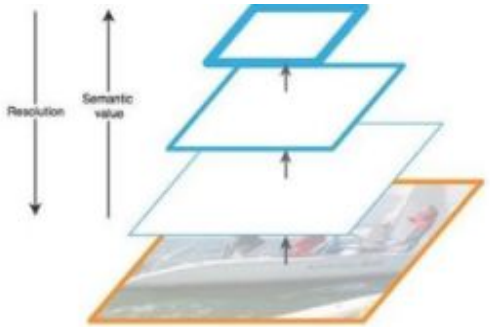


Figure 3.3: Top Down Pathway

SEQUEEZE AND EXCITATION NETWORK

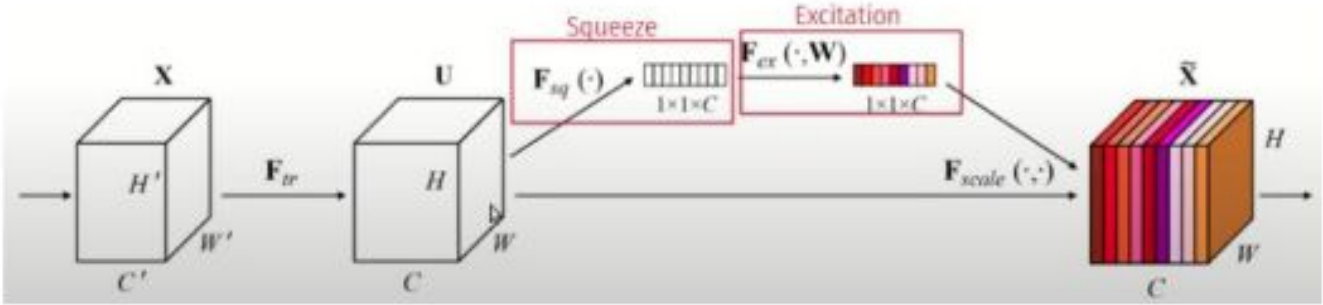


Figure 3.4: SE-Net architecture

SE-RESNET

To allow our model to emphasise important information features and suppress less useful features, we using Squeeze and Excitation Network(SE-Net) integrate bottom-up backbone ResNet101.

The architectural unit is designed to improve the representational power of a network by enabling it to perform dynamic channel-wise feature recalibration.

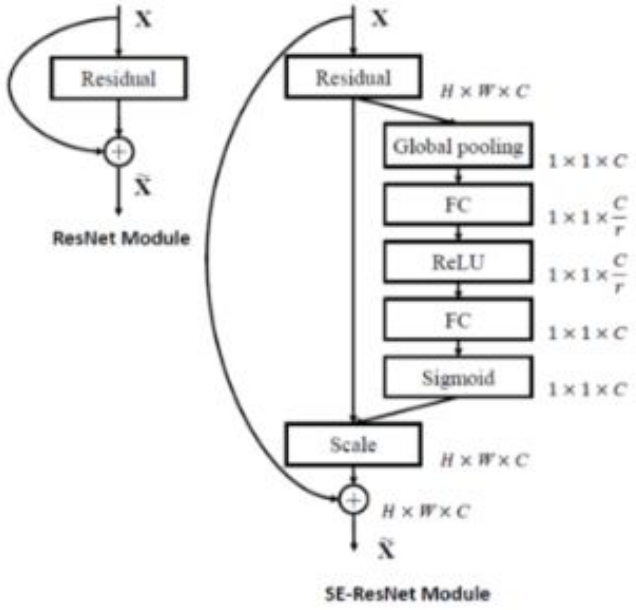
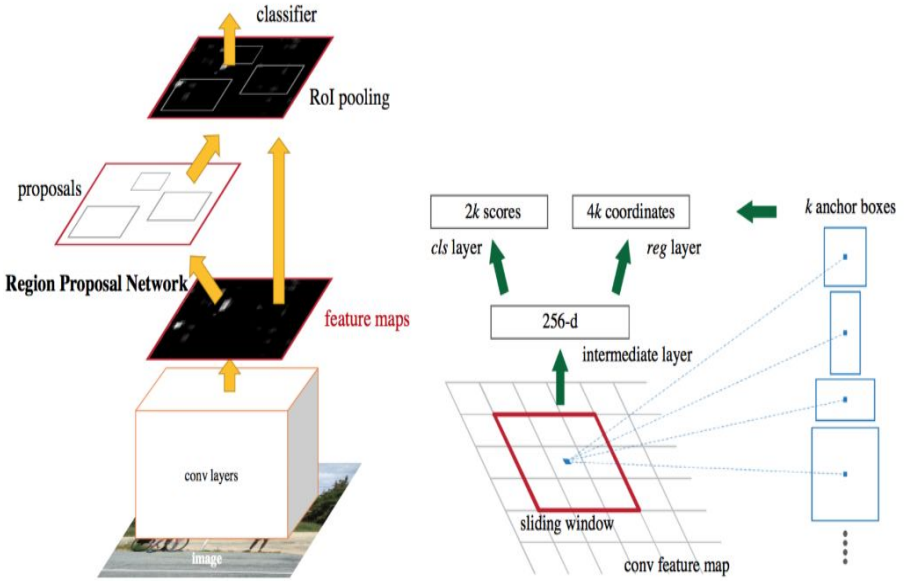


Figure 3.5: ResNet and SE-ResNet architecture

REGION PROPOSAL NETWORK(RPN)

The Region Proposal Network (RPN) runs a lightweight binary classifier on a lot of boxes (anchors) over the image and returns object/no-object scores.

Anchors with high objectness score (positive anchors) are passed to the stage two to be classified



REGION PROPOSAL CLASSIFY

After above step, all bounding box have transfer to Proposal classification stage. This stage takes the regional proposals from the RPN and classifies them

Region of interest before refinement

Region of interest after refinement

Region after filter low confidence and apply non max suppression

Figure 3.7: RPN classification for each stage

GENERATE MASK

This stage takes the detections (refined bounding boxes and class IDs) from the previous layer and runs the mask head to generate segmentation masks for every instance.

Figure 3.8: Target mask of dataset and prediction mask of mode

Figure 3.9: Final prediction of model

IMPLEMENTATION DETAIL

DATASET

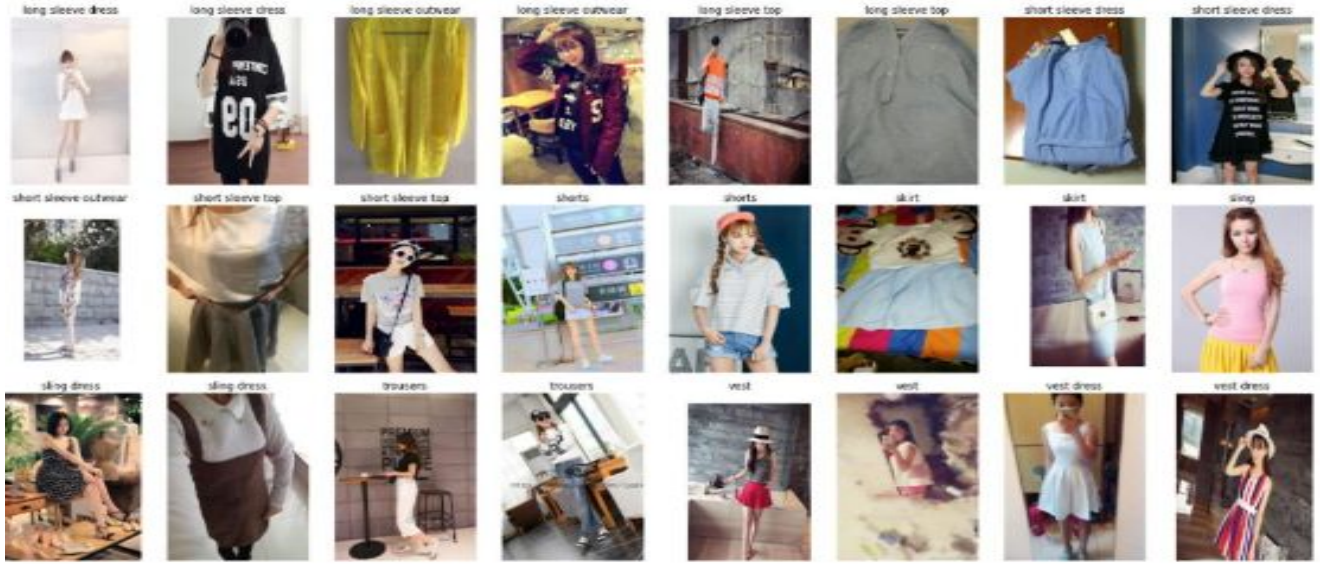


Figure 4.1: Example categories clothes image from our dataset

DeepFashion2 contain 491K photos from both commercial retailers and consumers of 13 popular clothing categories. The dataset is split into a training set (391K images), a validation set (34k images), and a test set (67k images)

STATISTICAL DATASET

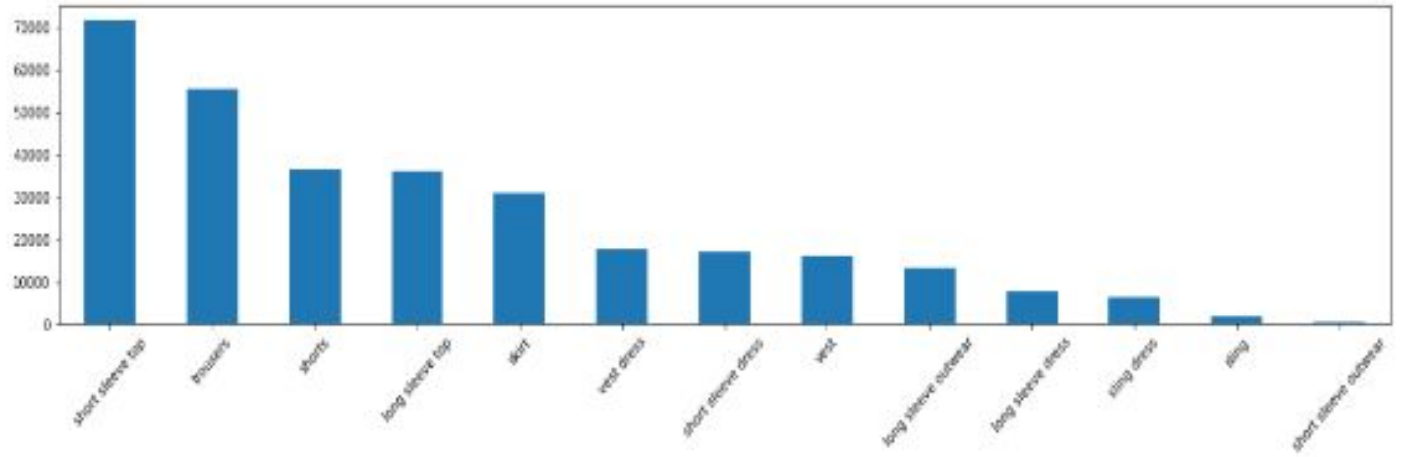


Figure 4.2: Statistic sample per categories of training set

DATA
PRE-PROCESSING

Data Argumentation with rotate images by 25 degrees and add noise to generate more data in classes with small sample sizes.

Figure 4.3: Example of our Data Argumentation

MINI MASK

Numpy uses 1 byte to store 1 bit value. Therefore, with an image size of 1024x1024, we need 1MB of ram memory to store it.

It quite large and consuming memory, make slow down the training speed of the model.

For improvement, instead of saving the entire mask of the whole image, this method only save the pixels of the mask in the bounding box.

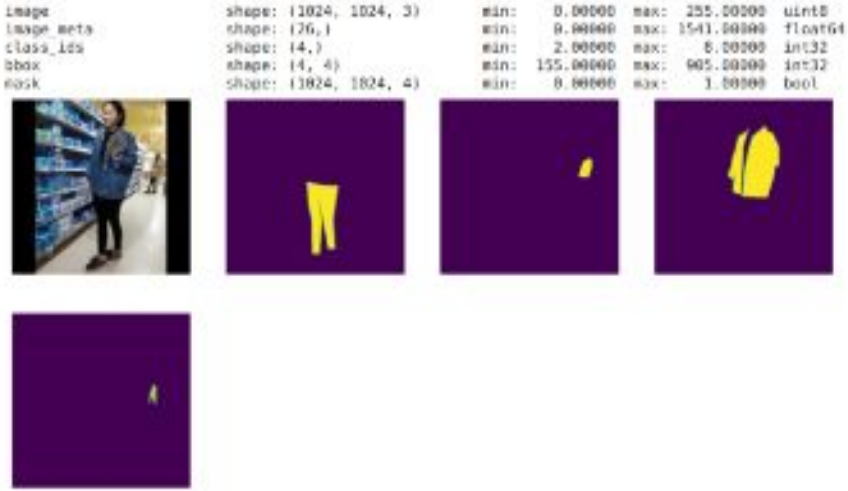


Figure 4.4: Mini mask

IMPLEMENTATION DETAIL

Pre-training with Common Object in Context(COCO) 2014 minimal(35k images)

Then load weight and use DeepFashion2 (limit 10000 samples per classes) to fine-tuning models, excluding the weight of fully connected layers(different number classes)

Learning rate start with 0,001 and down to 0.00025 from second stage. SGD optimizer with a weight decay of 0.0001 and momentum of 0.9.

Testing on NVIDIA T4, the training time is 20 days

EXPERIMENTAL AND RESULT

EVALUATION

Using DeepFashion2 validation(34k images) to evaluate model

Evaluation metric: Average Precision(AP)

QUALITATIVE RESULTS

Front view point

Side view point

Occlusion

Large Scale

Figure 5.1: Model results on different image conditions

Multiple pose

Figure 5.2: Model results on multiple human pose

QUANTITATIVE RESULTS

	APbox	APmask
AP	63.395	64.207
AP(IoU=0.50)	74.340	74.304
AP(IoU=0.75)	71.164	72.028
APsmall	30.050	26.386
APmedium	44.432	37.941
APlarge	63.613	64.531

AP box and mask of difference IoU and area(small, medium, large)

COMPARE WITH OTHER METHODS

Method	APbox
Match R-CNN	0.667
Aggreation and Finetuning	0.764
DeepMark	0.723
DeepMark++	0.737
Our	0.633

Compare AP box with other methods

	Aggreation and Finetuning	DeepMark++	Our
Short sleeve shirt	0.867	0.804	0.703
Long sleeve shirt	0.814	0.724	0.673
Short sleeve outwear	0.54	0.347	0.417
Long sleeve outwear	0.823	0.724	0.721
Vest	0.761	0.679	0.656
Sling	0.656	0.422	0.449
Shorts	0.784	0.721	0.629
Trousers	0.81	0.739	0.658
Skirt	0.818	0.74	0.752
Short sleeve dress	0.807	0.721	0.659
Long sleeve dress	0.659	0.542	0.54
Vest dress	0.812	0.71	0.704
Sling dress	0.773	0.605	0.675

Compare APbox of each class with other method

CONCLUSION AND FUTURE WORK

CONCLUSION

We proposes a lightweight method for fashion parsing tasks,can be integrated into any feature extraction backbone of detection and segmentation model for fashion parsing tasks by paying attention to important and less important features.

Mean Average Precision is used as our evaluation metric to compare the predicted bounding box and mask with the ground truth in the image of the dataset.

The approach did not give good results in testing. So we need to other method for solve it.

FAILURE CASES

Figure 6.2: Failure case with mask labelling

Figure 6.1: Failure case with detection

FUTURE WORK

Our work is also limited by the resolution of the training data, because training high resolution images require a large amount of computing power that currently is not available

The mask-rcnn model through our testing also shows that it does not give good results with the studied problem and we believe that the above results can be improved by simpler methods and less computational resources.

THANKS YOU FOR LISTENING