ISSN (e) 2520-7393
ISSN (p) 2521-5027
www.estirj.org

Fashion Image Analysis and Comparative Research Using Different Optimizers

Fatima Khanzada[1], Asif Aziz Memon[2], Rizwan Badar Baloch[3]

[1] *Department of Information Technology, Mehran University of Engineering and Technology*
[2]*Department of Computer Science, Dawood University of Engineering and Technology, Karachi*
[3] *Department of Computer System Engineering, Mehran University of Engineering and Technology*

Abstract: Fashion sector is one of the most important in the world, having a yearly revenue growth rate of 8.4%. Fashion image analysis proves to be more intricate than conventional image analysis due to the considerable diversity in styles, designs, and appearances. These variations frequently add complexity to the tasks of detecting and retrieving clothing items, making them challenging and intricate. One of the performance indicators for this research is accurate clothing detection; nevertheless, technological problems like the accessibility of huge datasets and the amount of time it takes to detect should be considered. The proposed network is a one-stage detector designed for rapid identification of diverse clothing items within the Kaggle (Fashion Product Images Dataset)**.** This suggested network enhances its backbone feature network by employing compound scaling and simultaneously training key input features at varying resolutions. It carefully balances the trade-off between interference time and precision through dedicated networks for bounding box prediction. Furthermore, it achieves efficiency gains by maintaining a low computational cost and a minimal parameter count. Through a comprehensive comparative analysis of diverse optimizers, encompassing Momentum, RMSProp, and Adam, it was determined that the Adam optimizer outperformed the rest, delivering a commendable accuracy rate of 96%. Remarkably, the trained model distinguishes itself by not only detecting single clothing items but also multiple garments within a single image. Furthermore, the model is incredibly lightweight and well-suited for use on low-power devices.

Keywords:*one-stage detector, Optimizers, low-power devices, Fashion image analysis,EfficientDet_D0*

1. Introduction

Fashion and clothing-related concerns are becoming more and more integral to individuals' everyday existence, the fashion sector has established itself as a substantial driver of the economy, particularly in the online sphere, where a substantial volume of fashion-related commerce is taking place. Customers are also displaying a keen interest in online shopping systems, enabling them to avoid the need to visit shopping malls and spend hours in search of their desired clothing items.

Existing datasets have a restricted number of annotations and struggle to address the many issues that arise in real-life applications due to abundance of extensive fashion datasets containing intricate annotations, the field of fashion image analysis has garnered considerable scientific interest. However, the deep learning models that is currently available for fashion datasets frequently have significant computing needs. In this research, we present a novel method designed specifically for energy-efficient devices.The single-stage detector that is suggested in this research can predict end-to-end clothing class categorization while also performing bounding box detection.

Object detection plays a pivotal role in fashion image analysis as it enables us to understand the content of an image and locate objects within it. There are two distinct approaches to object detection: the two-stage detector and the single-stage detector. It is crucial to carefully consider several important characteristics, such as localization accuracy, inference speed, and overall precision, while choosing the best detector to use. Region proposal approaches [1] and region-based convolutional neural networks (R-CNNs) [2] are key components of many object identification systems. The two-stage detector first recognizes region suggestions, then does categorization. The single-stage detector [3], [4] manages both region recommendations and classification at the same time, in contrast. For instance, localization accuracy and precision are exceptional in the two-stage detector Faster R-CNN [5]. On the other hand, the YOLO single-stage detector delivers quick inference speeds [3]. These factors show that the two-stage detector is a more flexible and precise choice than the single-stage detector. Single-stage detectors, on the other hand, are recognized for their effectiveness and quick inference.

Fashion plays a vital and multifaceted role in society, touching upon various aspects of our lives and culture. It serves as a powerful means of self-expression and identity, allowing individuals to communicate their values and personal style. Furthermore, fashion reflects and celebrates cultural diversity, preserving traditions and aesthetics economically, the fashion industry is a global powerhouse, contributing significantly to employment and GDP in many countries across manufacturing, retail, design, marketing, and modeling sectors.

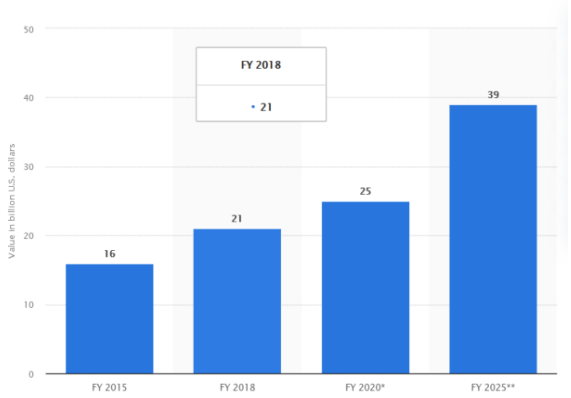

Figure 1 Market Value of Fashion in billion USD

Figure 1 shows the global value manufacturing, retail, design, marketing, and modeling sectors. Figure 1 shows the global value of fashion related industry. In 2015 it was close to 16 billion USD whereas it is projected to reach close to 40 billion USD till 2025 in just 10 years.

2. Related Work

This section provides an overview of the prior research and studies relevant to the method being proposed.

2.1. Fashion-Net

Fashion Net [6] as introduced by Liu in 2016, is a deep learning-based fashion image analysis system that utilizes the Deep Fashion dataset. This network is built upon the architecture of VGG-16 [7] and is tailored for the prediction of landmark positions and fashion attributes through modifications to its final layer.

Fashion Net made a significant contribution by utilizing the groundbreaking Deep Fashion dataset for fashion image analysis. However, one limitation of the Deep Fashion dataset is that it contains only a single clothing item. clothing analysis, which represents a notable constraint in its capabilities.

2.2. Match R-CNN

Ge proposed Match R-CNN [8] is a clothing detection network that makes use of the DeepFashion2 dataset. It conducts clothes identification, landmark estimation, and apparel segmentation. It is based on Mask R-CNN. The architecture comprises three interconnected networks: the Feature Network (FN), the Perception Network (PN), and the Matching Network (MN). FN is responsible for feature extraction and provides the PN with an output feature map using RoIAlign, a technique introduced in [39]. Utilizing the feature maps generated by FN, the PN carries out three primary tasks: landmark estimation, clothing detection, and segmentation. When it comes to bounding boxes and masks, its architecture has a similar arrangement that of Mask R-CNN. Landmark estimation and clothing segmentation both employ fully convolutional networks. MN, which is the Matching Network, serves as a crucial component in learning about similarities in clothes, effectively acting as a network for clothing retrieval. Research has significantly advanced thanks to Match R-CNN, which provides a useful guide for carrying out fashion picture analysis that includes a variety of garments [18]. However, because of its reliance on the Mask R-CNN's two-stage object identification structure, it has substantial inefficiencies in terms of inference performance and resource use. The actual implementation of fashion picture analysis in real-world contexts may be hampered by these limitations.

2.3. DFA Framework

A DFA approach was presented by Liu et al. [10] and uses a large fashion landmark dataset that includes more than 120k fashion photos. Each image in this collection has detailed annotations for eight different landmarks. Based on their real location and visible qualities, these photos are further divided into five groupings. Normal, medium, and big postures as well as medium and large zoom-ins are all included in these subgroups. It is important to note that these subgroups differ significantly in both the spatial and visual elements, with more than 30% of the photos exhibiting significant posture and zoom-in changes.

The DFA framework is divided into three phases to address these issues, with each level building on the predictions made in the stage before it.Consequently, DFA operates on entire images and delivers superior performance compared to other models like Deep Pose [11], all while demanding less computational resources.

3. Proposed Method

In our proposed method, we aim to develop a fashion image analysis network that effectively leverages the distinctive features of the fashion domain, ensuring its applicability in real-world scenarios. Our primary focus during the design process is to strike the right balance between accuracy and speed, while also optimizing resource utilization.
The primary objective of this study was to address the challenge of balancing accuracy and speed in fashion image analysis. The research successfully achieved an impressive inference time of just 39 milliseconds, demonstrating an optimal equilibrium between these factors when compared to existing methodologies. To achieve these objectives, we draw inspiration from Google Brain's Efficient Det [12], which has set the standard for state-of-the-art performance on the COCO dataset and is renowned for its ability to meet the conditions we seek. Notably, Efficient Det excels in terms of resource efficiency due to its low number of parameters, making it an efficient choice. Additionally, it offers fast inference speed and high accuracy.

In our adaptation of Efficient Det for the fashion domain, we have made global modifications to tailor the network to the specific requirements of fashion image analysis. This includes introducing a prediction head and designing a specialized loss function that aligns with our objectives in the fashion context. In the realm of fashion image analysis, it's imperative to consider the inherent features of the fashion domain to ensure practical scalability to real-world applications. In the design of our fashion image analysis network, we placed a strong emphasis on its applicability to real-world scenarios. Striking the right balance between accuracy and speed, along with resource efficiency, was a key objective. To meet these criteria, we turned to the Efficient Det model, initially introduced by Google Brain. Efficient Det has attained a state-of-the-art performance in

the COCO dataset and stands out as one of the most suitable models for these requirements. Its efficiency, characterized by a low parameter count, allows for resource conservation, while simultaneously offering fast inference times and high accuracy

3. Methodology

This section outlines the approach employed in this study to identify clothing items and describes how our system was put into action.

3.1 Data Collection

The research utilized a dataset that combines images from the Kaggle Fashion Product Images Dataset (available at https://www.kaggle.com/paramaggarwal/fashion-product-images-dataset)**[14]** where each product item is supplied with the image, title, and several attribute values of the product like Gender, Master Category, Sub-Category, Article Type, Base Color, Season, Year, Usage**,** Google, Freepik**[15]** and our own self taken images.

Table .1. Number of examples of each class

Label	**Quantity**
Bag	200
Belt	200
Dress	200
Glasses	200
Jeans	200
Nail-polish	200
Sari	200
Shirt	200
Shoes	200
Short	200
Sneaker	200
Socks	200
Tie	200
T-shirt	200
Watch	200
Wallet	200
Total	3200

3.2Dataset Labeling

In this step the collected dataset is labelled through the LabelImg tool

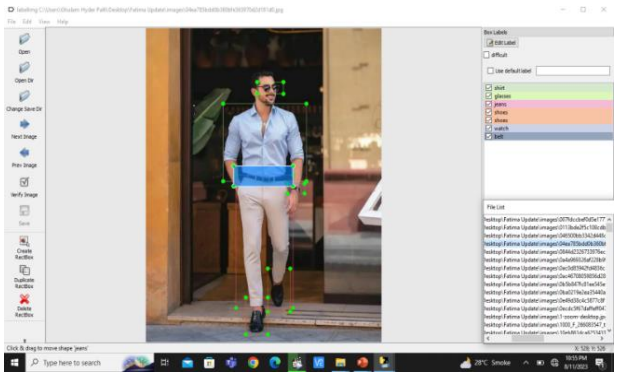

Figure 2 Image being labeled using LabelImg tool

xml files generated by lableImg tool are converted into csv files such as train_labels and test_labels to generate the tfrecord files.

3.3 Dataset Splitting

After the labeling of Dataset, it is divided into train, valid and test sets for model training and evolution. TF records are generated from these split data as shown in Figure.3.

Figure 3 Dataset classes split into train, test and validation

3.4 Model Training

The model is trained by TensorFlow 2 Object Detection API by using the efficientdet_d0.

EfficientDet [12] is a model designed to prioritize two tasks, ensuring it achieves the best possible trade-off between interference time and precision. In our specific context of fashion analysis, particularly in the realm of fashion landmark estimation, it's essential to initially assess whether EfficientDet is well-suited for this purpose.

Chen et al. [13] achieved exceptional results by employing a cascaded Feature Pyramid Network (FPN)to estimate critical points in images of different sizes. They introduced a subsidiary network called GlobalNet, which combines FPN and efficiently addresses the balance between high-quality and low-quality feature maps by performing element-wise addition following channel alignment through a 1 × 1 convolution operation

Furthermore, the MultiPoseNet [16] approach also demonstrated promising outcomes by harmonizing the number of channels following the passage of hierarchical Convolutional Neural Networks (CNN) through the FPN-based feature map. Consequently, there arises a necessity to adjust the influence of input features on the output features.

To address this requirement, we introduced a BiFPN structure that leverages cross-scale connections. This modification assigns additional weights to the input feature maps, thus allowing for a tailored adjustment of their contribution to the output feature maps. This strategic approach plays a significant role in enhancing the model's performance by enabling it to learn the importance of input features more effectively.

The input resolutions employed in our model are 512 × 512 and 640 × 640, and the backbone network is based on EfficientDet-D0. During training, we utilize a batch size of 16.

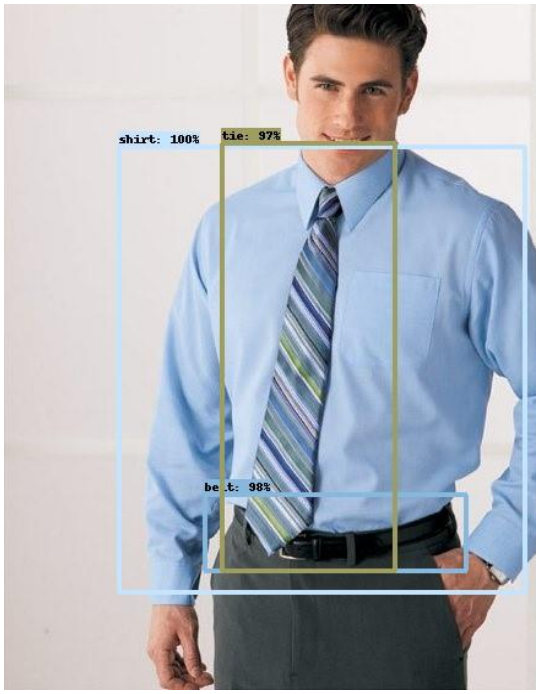

Figure 5 Multiple Fashion Detection using EfficientDet_D0

4. Results and Discussion

3.1 Intersection over Union (IoU):

The Intersection over Union [9] (IoU) measures the extent of overlap between a predicted bounding box and a ground-truth bounding box. The IoU loss quantifies the regions of intersection and union between these bounding boxes to determine the coordinates of their four respective vertices.

$$IoU = \frac{Area(B_p \cap B_{gt})}{Area\,(B_p \cup B_{gt})} \quad (1)$$

An IoU of 1 indicates a perfect match between two detected bounding boxes, while an IoU of 0 signifies that the two bounding boxes have no overlap or match. We set IoU (Threshold≥0.80).

Figure 5 shows a glimpse of IoU achieved by our model on different classes of the dataset.

Figure 4IoU on different classes of our dataset

Table 2IoUAcheived on different classes

S. No	Label	IoU (Threshold≥0.80)
1.	Bag	0.8903
2.	Belt	0.8640
3.	Dress	0.8777
4.	Glasses	0.9667
5.	Jeans	0.8993
6.	Nail-polish	0.9258
7.	Sari	0.9006
8.	Shirt	0.9200
9.	Shoes	0.8721
10	Short	0.9156
11.	Sneaker	0.9417
12.	Socks	0.9658
13.	Tie	0.9772
14.	T-shirt	0.9566
15.	Wallet	0.9459
16.	Watch	0.9547

*3.2*Comparative Research Using Different Optimizers:

The fashion detection is carried out by training the model using TensorFlow 2 Object Detection API. In this efficient-det (efficientdet_d0) is used with different optimizers [17] such as:

1. Momentum Optimizer

2. RMSProp Optimizer

3. Adam Optimizer

The main purpose of using different optimizers to train the model is to get the better accuracy so that the proposed research could benefit the fashion industry.

The reasonable accuracy is achieved by training the model on 16 fashion classes with different optimizers.

We performed experiment by applying Momentum, RMS_Prop and Adam to our model. Adam Optimizer is

found the best optimizer among the others with a reasonable accuracy of 96%.

5. Conclusion

We introduce a novel strategy by adapting the EfficientDet model to the domain of fashion image analysis. Our innovation centers on optimizing the computational efficiency and inference speed of EfficientDet, making it suitable for near-real-time fashion image analysis even on low-powered or standard devices. Our proposed network is designed for speed and efficiency, conducting comprehensive fashion image analysis through both classification and bounding box regression in a single forward pass, functioning as a single-stage solution. It's worth noting that our approach stands out from prior research that primarily concentrated on analyzing individual clothing items within the DeepFashion dataset, as we extend our analysis to encompass multiple clothing items simultaneously. This method has the potential to make significant contributions to future endeavors in the field of fashion image research.

References

[1] R. Mo and Y. C[1] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, ‘‘Selective search for object recognition,’’ Int. J. Comput. Vis., vol. 104, no. 2, pp. 154–171, Sep. 2013.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ‘‘Rich feature hierarchies for accurate object detection and semantic segmentation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ‘‘You only look once: Unified, real-time object detection,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.

[4]W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, ‘‘SSD: Single shot multibox detector,’’ in Proc. Eur. Conf. Comput. Vis., Cham, Switzerland: Springer, 2016, pp 21–37.

[5] S. Ren, ‘‘Faster R-CNN: Towards real-time object detection with region proposal networks,’’ in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 1–9..

[6]Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, ‘‘DeepFashion: Powering robust clothes recognition and retrieval with rich annotations,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1096–1104.

[7] Mogan, Jashila Nair, et al. "VGG16-MLP: gait recognition with fine-tuned VGG-16 and multilayer perceptron." Applied Sciences 12.15 (2022): 7639.

[8] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, ‘‘DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and reidentification of clothing images,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5337–5345.

[9] Rahman, Md Atiqur, and Yang Wang. "Optimizing intersection-over-union in deep neural networks for image segmentation." In *International symposium on visual computing*, pp. 234-244. Springer, Cham, 2016.

[10] Z. Liu, ‘‘Fashion landmark detection in the wild,’’ in Proc. Eur. Conf. Comput. Vis., Cham, Switzerland: Springer, 2016, pp. 229–245

[11] A. Toshev and C. Szegedy, ‘‘DeepPose: Human pose estimation via deep neural networks,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1653–1660

[12] M. Tan, R. Pang, and Q. V. Le, ‘‘EfficientDet: Scalable and efficient object detection,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10781–10790

[13] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, ‘‘Cascaded pyramid network for multi-person pose estimation,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7103–7112

[14] “Fashion Product Images Dataset | Kaggle.” https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset (accessed Aug. 13, 2023).

[15] “Freepik: Download Free Videos, Vectors, Photos, and PSD.” https://www.freepik.com/ (accessed Aug. 17, 2023).

[16] M. Kocabas, S. Karagoz, and E. Akbas, ‘‘Multiposenet: Fast multi-person pose estimation using pose residual network,’’ in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 417–433.

[17] “tensorflow - Difference between RMSProp with momentum and Adam Optimizers - Data Science Stack Exchange.” https://datascience.stackexchange.com/questions/26792/difference-between-rmsprop-with-momentum-and-adam-optimizers (accessed Aug. 12, 2023).

[18] Kim, Hyo Jin, et al. "Multiple-clothing detection and fashion landmark estimation using a single-stage detector." IEEE Access 9 (2021): 11694-11704.hew, "MMSE-based joint source and relay precoding design for amplify-and-forward MIMO relay networks," *Wireless Communications, IEEE Transactions on*, vol. 8, pp. 4668-4676, 2009.

Author 1: FATIMA KHANZAD received the B.E degree from the Mehran University of Engineering and Technology (MUET), Jamshoro, Pakistan, in 2018 respectively currently doing M.E from the Mehran University of Engineering and Technology (MUET), Jamshoro, and working as Senior Bidding Coordinator in Basecamp Data Solution. My research interests include machine learning, image processing and Deep learning.

Author 2: ASIF AZIZ MEMON received the B.E. and M.E. degrees from the Mehran University of Engineering and Technology (MUET), Jamshoro, Pakistan, in 2010 and 2015, respectively. He received his Ph.D. degree in Application Software from Chung-Ang University, Seoul South Korea, in 2021. Currently, He is working as an Assistant Professor and Head of the Cyber Security Department at Dawood University of Engineering and Technology, Karachi. His research interests include image segmentation, image recognition, and medical imaging.

Author 3: RizwanBadarBaloch received the B.Eand M.E. degrees from the Mehran University of Engineering and Technology (MUET), Jamshoro, Pakistan, in 2004 and 2008, respectively.He is working as an Assistant Professor in Mehran University of Engineering and Technology. His research interests include Object Oriented Programming and Machine Learning.