Image Classification and Visual Search for Fashion Items Using Deep Learning Algorithms

Agboola Tawakalit Omolabake
Department of Computing and Mathematics.
Manchester Metropolitan University.
Manchester, United Kingdom.
23653040@stu.mmu.ac.uk

***Abstract*—Manual tagging of fashion product images affects the efficiency of e-commerce platforms, as well as a lack of an efficient search function to manage efficiency. In order to address the problem, fashion product images require accurate classification and a search model to yield satisfactory results. An extreme solution of applying deep learning algorithms would present the use of VGG19 to correctly tag images using artificial labels and CNN autoencoders to retrieve items with similar images. This paper determines that the VGG19 model has a classification accuracy of 92.5%, and the visual search in the autoencoder configuration can also retrieve similar images. Also, the other deep learning models' classification and performance of visual search were the initial basis for comparing with our results, indicating an improvement in efficiency. This paper demonstrates that enhancing e-commerce operational efficiency can be achieved by using identified image classification.**

***Index Terms*—Deep Learning, Convolutional Neural Networks, Image Classification, Visual Search, VGG19, RestNet50, Autoencoders, Data Augmentation, E-commerce.**

I. Introduction

E-commerce platforms change the way we shop and help choose between thousands of products in minutes. However, many challenges remain, with product classification and visual search being some of the pressing issues. Sellers struggle to tag products accurately, leading to search and visibility problems, and buyers find it hard to search for products with extensive unfamiliarity with the most precise queries. This paper presents the use of deep learning solutions to address these problems and improve the seller and buyer experience.

Another challenge is that the demand of a customer is suffering because he does not know proper keywords. A buyer must enter the keywords through which the products matching with the keywords must appear on his display. If a customer knows the product properly, they can easily keyword search and making an order. But customers are not necessarily knowledgeable about every other product. Visual search is a solution to this problem. Visual Search is a product search by photo or image. Customers can take a photo of a product send the search engine a picture of the product through the website and the visual search engine interprets the image and provides the visually similar products that a shopper can buy. Many AI solution providers are using this idea such as Visenze, Google Lens, etc. Any visual search algorithm can be an unsupervised problem where machine learning models learn the features of the new images and search for the similar products by their feature. Once the target image is uploaded to the visual search, the same features can be used to render more products on the website. Standard visual search algorithm such as autoencoders can also generate the latent features of images.

This research has been initiated with the goal of addressing these limitations, where using deep learning methods are leveraged to increase the accuracy of image classification and the efficiency of the visual search algorithm. In particular, two aims have been identified for this purpose, which includes;

- The enhancement of product tagging accuracy with the employment of image classification based on the VGG19 model.
- The development of an autoencoder for intuitive visual search.

II. Literature Review

Deep learning has significantly assisted in development of e-commerce through improving image recognition, which helps in product categorization and recommendations, among other aspects of it. There have been several approaches to address the problem of image classification and visual search in the e-commerce domain. In particular, deep learning models, such as convolutional neural networks can learn hierarchical representations from raw pixel data and are, therefore, commonly used in this context. With regards to image classification, AlexNet is one of the pioneering CNN models that outperformed other competitors in the ImageNet Large Scale Visual Recognition Challenge[9]. AlexNet consisted of layers of convolutions, pooling, and fully connected layers and used data augmentation and dropout to reduce overfitting. Meanwhile, InceptionV3 introduced several innovations of architecture, including inception modules which were more computationally efficient due to better utilization of model parameters, and it achieved state-of-the-art performance on various image classification benchmarks[13]. Finally, ResNet solved the problem of vanishing gradients in deep networks by using residual blocks, allowing them to be trained much deeper and with more accuracy.

Autoencoders have also been utilized to learn a compact representation of an image, which can later be used to find similar images based on the learned features. Finally, hybrid models based on combining different models have been proposed. Indeed, all of the methods have certain strengths and limitations. For instance, AlexNet and InceptionV3 score well on accuracy, but in e-commerce applications, the data is much more diverse and unstructured than the dataset on which these models were trained. ResNet models, with their residual connections, have significantly improved the state of the art, but in real-time and resource-constrained environments, what matters is their production-level implementation, and this remains largely unsolved[4]. Regarding image visual search, deep metric learning and auto-encoders seem to be very promising, but the balance must be found between high scores and computational costs when the dataset becomes very large[11]. Most of the existing methods outperform other known solutions with regards to the academic benchmarks, and the potential and room for improvement remain for practical use.

III. DATASET

A. Dataset selection

The dataset used in the present study was taken from Kaggle[8], a well-known platform for machine learning datasets. In particular, it contains a broad spectrum of fashion product images, ensuring great coverage of diverse categories. This dataset is especially valuable, as it represents the variety and richness of E-commerce product image data in the real world. The dataset includes images of clothes, footwear, accessories, as well as metadata, including product identifiers, categories, and attributes. There are thousands of images in the dataset, allowing for sufficient training and testing data volumes. Different categories of products and their styles enable the creation of a reliable classification and visual search model.

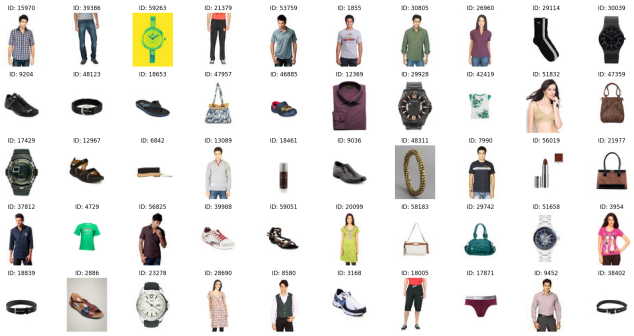

Fig. 1: Random selections of Fashion Product from the Dataset

B. Dataset Integration

The sets of images and accompanying metadata are imported into a Google Collaboratory environemnt. The decision was made to employ the Google Collaboratory because it has simple and ready access to a range of powerful computational resources without an additional financial burden, which include GPU and TPU. This environment is used for adept handling and processing of massive datasets, which is especially important when it comes to processing highly computationally demanding deep-learning models.

C. Data Preparation and Preprocessing

An initial analysis of the dataset was performed to understand various key characteristics. These include the image dimensions, color distribution, and the proportion of each category of products. Knowing these properties is essential for customizing pre-processing steps to ensure the deep learning models operate at optimal conditions.

- Normalization: The pixel values of the images were normalized to the range [0, 1]. Normalizing the streamlines the pixel values, a factor that stabilizes and hasten the learning process of neural networks.
- Data Augmentation: Data augmentation techniques, including rotation, flipping, and zooming, were employed to artificially boost the size of the training dataset, enhancing the model's generalization capabilities. Similarly, data augmented the images helps to familiarize the model with different conditions and angle viewpoints
- Resizing The images were resized to the uniform dimensional of 224×224, a format suitable for feeding to the VGG19 and the Autoencoder models. The resizing step ensures uniformity and a compatible format for neural networks to express calculations.

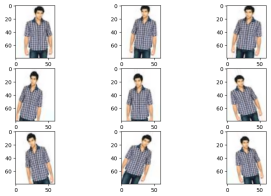

Fig. 2: Sample Image After Augmenting the Dataset

IV. METHODOLOGY

This research's methodology is best described as an overarching approach based on advanced deep learning, customized to work on e-commerce efficiency through image classification and visual searches. The research at hand applied a hybrid deep learning architecture, combining VGG19 for image classification with a customized version for fashion, and utilized autoencoders and ResNet-50 for the visual search capabilities[12]. The research's originality and prior contribution lie in the specified use of the models based on the specificity of fashion images. The architecture integrated VGG19 due to a deep structure well suited for scaling operations and dealing with small differences in fashion categories. Autoencoding was incorporated to reduce the dimensions of the images and perform feature extraction[5], which is crucial for later visual explanations. A high-level explanation of the presented diagram is sufficient to understand the captured workflow.

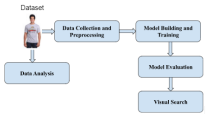

Fig. 3: Design of Experiment

A detailed diagram demonstrates the flow from image input to the several processing stages to final outputs in the form of classified categories and visual search results. To improve the overall system's performance both in accuracy and time, this configuration utilizes different deep learning architectures designed for specific tasks.

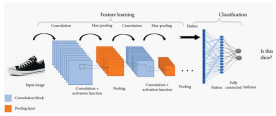

Fig. 4: Illustration of the working of CNN

A. *Learning Algorithm*

This research introduces a hybrid approach combining the strengths of VGG19 for classification and CNN autoencoders enhanced with ResNet-50 for visual search. This dual methodology improves accuracy and efficiency in handling fashion product images, making it a robust solution for e-commerce applications.

- VGG-19: VGG19 is a deep CNN architecture with 19 layers, including 16 convolutional layers followed by three fully connected layers and relatively few parameters remain in the fully connected layers[12].
- Convolutional Neural Networks (CNN): CNNs are a type of deep neural network for image processing. They have convolutional layers, which automatically identify spatial hierarchies of features based on prior experience, and independent of the unsupervised learning display program, the filters determine the

patterns of the input image 11. Feature patterns such as edges, textures, and patterns can be computed using them to make an attention map.
The output of a convolutional layer l can be expressed as:

$$h_{l+1} = f(W_l * h_l + b_l)$$

where h_l is the input feature map, W_l are the weights, b_l is the bias, $*$ denotes convolution, and f is the activation function.

- CNN-based Autoencoder: Autoencoders are a class of artificial neural networks that are used to gain efficient representations of data. As the name suggests, these networks are composed of two parts an encoder and a decoder. In the Encoder Part, the image is compressed by the encoder into a latent space representation. To decrease files' spatial dimensions and extract features, several convolutional and maximum pooling layers were employed. While the decoder reconstructs an image from a latent space representation. The-encoded characteristics were used to create the image using transposed convolutions and upsampling.
  The loss function for training an autoencoder is typically the mean squared error (MSE) between the input image x and its reconstruction $\hat{x}$:

$$L(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$$

- ResNet-based Autoencoder: ResNet-50 is an instance of the ResNe model with 50 layers. The residual blocks help to reduce the vanishing gradient problem and enable the training of very deep networks . Residual blocks allow the model to learn identity mappings that ensure the gradient is well-preserved during backpropagation[4]. As a result, the model can be made deeper and can learn more generic or complex image transitions in the feature space. The encoder part of an autoencoder of a ResNet-50 is used to extract image features, and the similarity metrics are used to compare the results. To extract the sixteenth-dimensional image features from the dataset for the visual search tasks, the trained encoder portion of the ResNet-50 is employed.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. Training, Validation and Testing Strategy

The experiment was designed and conducted to measure the performance for image classification using the VGG19 model and then tested the performance for visual search using a CNN autoencoder with Resnet-50. The dataset was divided into a 80:20 training and test ratio respectively, and the training set was splited into 80:20 training and validation ratios respectively for model building, [Training data: 3200, Validation data: 800 and Testing data: 1000].

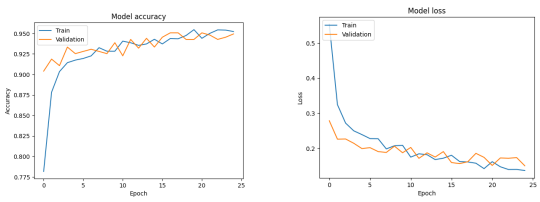

Fig. 5: Performance of VGG19 During Training

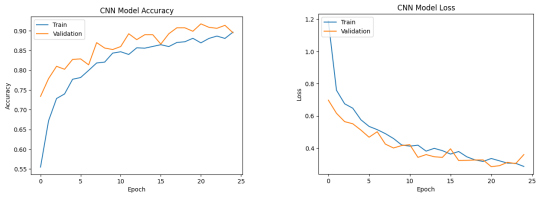

Fig. 6: Performance of CNN During Training

B. Confusion Matrix

Confusion matrix provides a detailed perspective on classification accuracy according to the product category, so one can figure out which product categories are often misclassified. The confusion matrix below visualizes the performance of VGG19 across different categories. This matrix helps in identifying the misclassification patterns and the categories with lower performance.

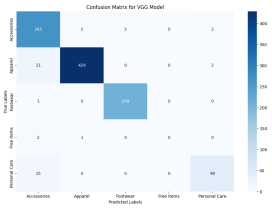

Fig. 7: Confusion Matrix of the Predictions made by VGG19

C. AUC Curves

The AUC curve for the classification task indicates a high performance with an AUC score of 0.95. The AUC curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) across different threshold settings. The precision-recall curve helps look at the overall performance of the visual search model paying attention to the balance between precision and recall.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

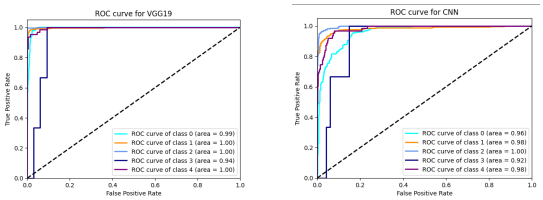

Fig. 8: ROC-AUC CURVE for VGG19 and CNN

D. Quantitative and Qualitative Results

The quantitative results indicate that the performance of the VGG19 model in the classification of fashion products' images is high. The confusion matrix in Figure 6 shows that most categories are

correctly classified. There exist confusion only between more visually similar categories and are sparse. The confusion matrix outlines the prediction of cases across the categories and shows which areas the model performed well and poorly. It has a clear indication of how well the model correctly predicts the matrix.
The visual search model using CNN autoencoders and Resnet-50 had a good performance model. It is capable of being used to retrieve visually similar items to the one being searched for. The results are summarized in tables and visualizations below:

TABLE I: Model Performance Comparison

Metric	VGG19	CNN Autoencoder	ResNet-50 Autoencoder
Accuracy	92.5%	85.0%	88.5%
Precision	91.8%	84.2%	87.7%
Recall	90.4%	83.6%	86.9%
AUC	0.95	0.87	0.89

E. Visual Search Performance:

The visual search model was tested by querying with various images. The top 5 similar images retrieved for a sample query image are shown below. This demonstrates the model's ability to identify visually similar products from the dataset. As shown from the visual search results in Figure 9, the autoencoder-based model has high practicality in retrieving visually similar objects. Thus, the qualitative analysis confirms the usefulness of the model in real e-commerce applications, as it allows users to find similar products based on visual similarity.

Fig. 9: Image Prediction Using VGG19 Model

Fig. 10: Virtual Search Using Autoencoder

F. Discussion of Result

From the experimental results, it could be inferred that the VGG19 model was highly accurate with a 92.5 % classification accuracy and a very high AUC score of 0.95. This observation points to the fact that this model was able to capture fine-grained features that are critical to classifying fashion images. As proven by the precision and recall rates, 91.8 % and 90.4 %, The model was able to classify a wide range of fashion products accurately. From the confusion matrix, it is observed that there were categories such as 'footwear' which had lower recall; the misclassification in this category happened more. This shows that the model performs well overall but experiences difficulty in fine distinction and might misclassify similar items; initially, overfitting was a concern only curbed through dropout and data augmentation validation techniques, though the precision, recall, and performance variations between test and train data indicates further improvement of the model.

The visual search model that combines the CNN autoencoders with ResNet 50 presented good performance in the retrieval of similar visual results, as evidenced by the qualitative results. In various queries, the model was able to identify top similar items, implying real-life applicability and user experience in e-commerce platforms since users can quickly search for and find similar items they like instead of spending time searching through thousands of items. Although the model showed good performance in the current problem, some categories were critical for further optimization to improve their performance.

G. Research Limitations

While the results may be promising, the study also has limitations. The dataset involved was sizable, it did not fully encompass the vast collection of different fashion products sold every day. In addition to not having a wide range of different items to recognize, the performance was negatively impacted by variations in picture quality and level of annotation. Utilization of VGG19 and ResNet-50 required significant computational resources and time, making it challenging to scale. In addition, the speed of inference is not sufficient for real-time applications, and the quality of generalization is not great. Overfitting was observed to some extent, implying that the model must become generalized to unknown samples. The model struggled to recognize the subtle distinctions between similar goods and used the input of non-visual information for additional interpretation. These limitation issues included overfitting and other challenges to generalization to new categories. This weakness can be overcome by advanced data augmentation, transfer study, and real-time prediction for the user's loop feedback to improve the models. Future work can be focused on accelerate the use of these models in real-time for improving the efficiency of the scalability and e-commerce and continually improving its prediction ability to remain relevant to the constant changes in the fashion industry.

VI. CONCLUSION

This research was conducted to improve the efficiency of e-commerce by improving the classification and visual search of fashion products' images utilizing learnings from deep learning. The research utilized the VGG19 model for classification which achieved high accuracy, precision, recall, and an impressive AUC score. The results, therefore, demonstrate that the VGG19 model is capable of capturing the small features that are highly useful in classification. Furthermore, a CNN autoencoder integrated with ResNet-50 was used for visual search which retrieved visually similar images proving practical approach to enhancing the use's convenience of e-commerce platform.

VII. PROTOTYPE

The link below is a supporting document that contain the technical part of the research.

https://github.com/TwiTech/Image-Classification-and-Visual-Search-Using-Deep-Learning-Algorithms/tree/main

REFERENCES

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171-4186.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.

[3] J. Gao, J. Fan, W. Jiang, and J. Han, "Hybrid Neural Networks for Learning the Trend in Time Series," in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017, pp. 1977-1983.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[5] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," Science, vol. 313, no. 5786, pp. 504-507, 2006.

[6] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861, 2017.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700-4708.

[8] Kaggle: Fashion Product Images Dataset https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems, 2012, pp. 1097-1105.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proceedings of the International Conference on Machine Learning, 2021, pp. 8748-8763.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815-823.

[12] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in International Conference on Learning Representations (ICLR), 2015.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818-2826.

[14] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 6105-6114.

[15] N. C. Mithun, R. Panda, and A. K. Roy-Chowdh ry, "Co str ctio of Diverse ma e Datasets from Web Collectio s with Limited Labeli ," IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 4, pp. 1147–1161, 2020, doi: 10.1109/TCSVT.2019.2898899.

[16] W. Zho , H. Li, a d Q. ia , "Rece t Adva ce i Co te t-based Image Retrieval: A Literature

[17] S rvey," pp. 1–22, 2017, [Online]. Available: http://arxiv.org/abs/1706.06064.

[18] J.-C. Chen and C.-F. Li , "Visual-based Deep Learning for Clothing from Large Database," dl.acm.org, vol. 07-09-October, Oct. 2015, doi: 10.1145/2818869.2818902.

[19] N. Khosla a d V. Ve katarama , "B ildi ma e-Based Shoe Search Using Convolutional Neural Networks." Accessed: J . 04, 2021. [O li e]. Available:http://vision.stanford.edu/teaching/cs231n/ reports/2015/pdfs/nealk_final_report.pdf

[20] Daniele Micci-Barreca. 2001. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. ACM SIGKDD Explorations Newsletter 3, 1 (2001), 27–32. Google ScholarDigital Library

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2018. Efficient estimation of word representations in vector space. Workshop in International Conference on Learning Representations (ICLR) (2018).Google Scholar

[22] Rishabh Misra, Mengting Wan, and Julian McAuley. 2018. Decomposing fit semantics for product size recommendation in metric spaces. In Proceedings of the 12th ACM Conference on Recommender Systems. ACM, 422–426. Google ScholarDigital Library

[23] Fanke Peng and Al-Sayegh. Mouhannad. 2014. Personalised Size Recommendation for Online Fashion. In 6th International conference on mass customization and personalization in Central Europe.

[24] Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV), Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.

[25] Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). Comput. Vis. Image Underst. 2008, 110, 346–359. [CrossRef]

[26] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), Columbus, OH, USA, 24–27 June 2014; pp. 580–587.

[27] Girshick, R. Fast R-CNN. 2015; In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 1440–1448.

[28] Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.

[29] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv 2015, arXiv:1506.01497

[30] Bianco, S.; Buzzelli, M.; Mazzini, D.; Schettini, R. Deep learning for logo recognition. Neurocomputing 2017, 245, 23–30