Fashion Item Classification using Deep Learning

<Jiahao Zhou (z5567496), Ling Tang(z5545055), Runxuan Tian (z5527073), Yu Peng (z5638271), Yuntian Zhang(z5675701)>

***Abstract*—This project explored the application of deep learning methods for the task of fashion product classification. Taking AlexNet as the baseline, we first conducted systematic ablation experiments by adjusting the number of convolutional layers and replacing classifiers (Softmax and SVM), and then introduced the CBAM attention mechanism module to optimize the performance of AlexNet and ResNet18 architectures. The experimental results based on the manually annotated fashion product dataset show that increasing the convolution depth and using the SVM classifier can improve the classification performance, while the introduction of the CBAM module can significantly improve the recognition ability of fine-grained categories such as accessories. Finally, the ResNet18 model combined with CBAM achieved the best results, with a macro-average F1 score of 0.93. This paper provides effective model design and iterative optimization ideas for the practical application of fashion product classification.**

Keywords—Convolutional Neural Networks (CNN), Deep learning, AlexNet, ResNet, CBAM (Channel and Spatial Attention), Fashion Classification

I. INTRODUCTION

With the rapid development of e-commerce, the number of pictures of clothing and fashion products on the Internet has increased exponentially [1]. Online shopping platforms need to quickly and accurately identify product categories through pictures to achieve automatic labeling, intelligent recommendations, inventory management, and personalized marketing [2]. Accurate classification of fashion items can help consumers shorten their search time, improve their shopping experience, and promote conversion rates [3].

Due to the diversity of design, fashion product images are highly complex and contain a large number of variants. Traditional methods are difficult to accurately identify. Deep Learning, especially CNNs, has demonstrated excellent performance in the field of image recognition and provided a solution for complex visual feature extraction [4]. Research on Fashion Item Classification can explore and promote the application of deep learning in scenarios such as fine-grained classification, feature extraction, and multi-label recognition. Specifically, the study of fashion item classification can explore and promote the application of deep learning in scenarios such as fine-grained classification, feature extraction, and multi-label recognition. Based on this background, the use of deep learning for fashion product image classification has attracted increasing attention from academia and industry. This not only helps promote the application of artificial intelligence(AI) in actual business scenarios, but also provides theoretical and technical support for building smarter and more efficient recommendation and image search systems.

This paper explores the potential of the mainstream CNN architecture in fashion item classification and continuously optimizes the classification ability of the model through fine-tuning. Specifically, this paper is divided into three studies and one pilot study. In the pilot study, we discussed and selected appropriate data augmentation methods and epochs, which will be used in the next three studies. In study 1, the methods of reducing or increasing convolutional layers and modifying classifiers were compared. In study 2, Channel and Spatial Attention (CBAM) was introduced into AlexNet and compared. Based on the conclusions of study 1 and 2, study 3 introduced a deeper ResNet architecture than AlexNet and combined it with CBAM to test whether it would have better performance. All in all, this paper provides an excellent model and iterative methodology for fashion item classification through three progressive studies.

II. RELATED WORK

Automatically classifying clothing images into standardized fashion categories is a fundamental challenge in computer vision. Deep learning, especially convolutional neural networks (CNNs), has become the paradigm of choice for this task due to its powerful feature extraction and pattern recognition capabilities. Early classic studies laid the foundation for the study of deep learning in the fashion field. Liu et al. (2016) established the large-scale fashion dataset DeepFashion, containing over 800,000 clothing images with 50 fine-grained categories. They proposed FashionNet, a model based on VGG-16, which improved feature extraction by predicting clothing landmarks to adapt to deformation, occlusion, and varying camera angles [5]. Corbière et al. (2017) employed ResNet-50 as the backbone CNN, initialized with ImageNet pre-trained weights for feature extraction. They utilized weakly annotated text-image pairs to address labeling challenges in e-commerce scenarios [6]. Zou et al. (2022) introduced A100, the first systematic framework to evaluate the aesthetic ability of fashion compatibility models, including LAT (Liberalism Aesthetic Test) and AAT (Academicism Aesthetic Test). This study offers cross-domain insights for clothing classification tasks, particularly in feature engineering and domain knowledge integration [7].

Reviewing the latest research, most works adopt the standard CNN architecture, usually using transfer learning from ImageNet pre-trained backbone networks (such as ResNet, EfficientNet, DenseNet, and VGG) to improve the performance of fashion category datasets [4][8][9][10]. Moreover, most studies focus on the Fashion-MNIST dataset, but a few studies extend their evaluation to more complex and diverse datasets, such as Street-FashionData[9][11]. As more and more niche brands use diversified channels such as independent websites for sales, lightweight fashion classification models are equally important. Based on this situation, this paper uses the dataset of Zou et al. (2022) for manual annotation to carry out the following research.

III. METHODS

This project uses convolutional neural networks (CNNs) for fashion product classification. The classic deep CNN model AlexNet is used as a baseline. First, ablation experiments that modify the number of convolutional layers and classifiers are performed (study 1), and as the experiments progress, the

CBAM attention module (study 2) and the ResNet architecture (study 3) are gradually introduced.

3.1 AlexNet

3.1.1 AlexNet (Baseline models)

It is a CNN architecture developed from the original AlexNet architecture [12], with an input image size of 224×224. AlexNet with 8 layers of depth is the winner of ILSVRC 2012 competition. The first 5 layers are convolutional, and the last 3 layers are fully connected. There are also activation and pooling layers among the layers. It has an important place among CNN models. Classic AlexNet architecture is presented in Fig. 1.

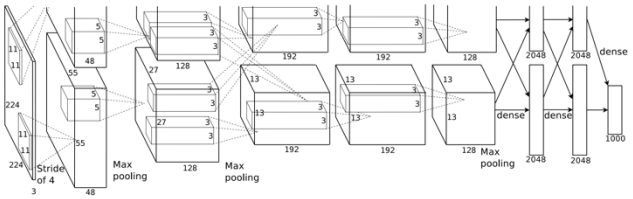

Fig. 1. AlexNet Architecture [12]

3.1.2 Proposed models of AlexNet with variations

3.1.2.1 AlexNet (change the number of convolutional layers or classifiers)

This ablation experiment is to verify the classification ability of AlexNet on fashion projects by changing the number of convolutional layers or classifiers of the AlexNet architecture, and to find potential better model architectures through fine-tuning. Specifically, referring to the study of Eldem et al. (2023), a series of ablation experiments were designed[13]. There were 7 groups in total, and the number of convolutional layers (3 layers, 4 layers, 6 layers) and the type of classifier (Softmax and SVM) were adjusted respectively. Among them, Study 1 adopted the standard AlexNet structure, including 5 convolutional layers and Softmax classifier; Study 2 to Study 7 made systematic changes in the depth of the convolutional layer and the final classification module to examine the impact of different structural combinations on classification performance (Table 1). By comparing the classification effects of each model, this study aims to analyze the specific mechanism of the number of convolutional layers and classifier selection on the accuracy of fashion product image recognition.

Table 1

Ablation experiment of AlexNet

Group	Model	Conv layer	Classifier
1	AlexNet	5	Softmax
2	3Conv_Softmax	3	Softmax
3	3Conv_SVM	3	SVM
4	4Conv_Softmax	4	Softmax
5	4Conv_SVM	4	SVM
6	6Conv_Softmax	6	Softmax
7	6Conv_SVM	6	SVM

3.1.2.2 AlexNet + CBAM (Channel and Spatial Attention)

In this improved methodology, we add CBAM Attention Module to the AlexNet model to increase the model's performance on the Accessories low recognition accuracy optimization.

CBAM is a lightweight, general-purpose attention module that contains two parts of the attention mechanism: Channel Attention and Spatial Attention. Classic CBAM architecture is presented in Fig. 2.

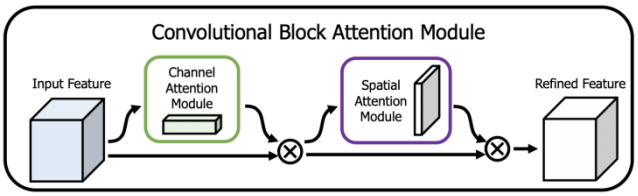

Fig. 2. CBAM Architecture [15]

According to the reference paper[14], we decided to add the attention mechanism in layers 3 to 5. Three new models were proposed by inserting the CBAM attention module with different parameters

Table 2

Parameters of CBAM Attention(Ratio, Kernel_size)

Model	Ratio	Kernel_size
AlexNet_CBAM(16&7)	16	7
AlexNet_CBAM(8&7)	8	7
AlexNet_CBAM(8&5)	8	5

3.2 ResNet

3.2.1 ResNet (Baseline models)

Our baseline model uses the pretrained ResNet18 architecture with all convolutional layers frozen. ResNet is a deep convolutional neural network structure proposed by Microsoft Research in 2015, which mitigates the problem of gradient vanishing and degradation in deep neural networks by introducing the 'Residual Connection'. ResNet architecture is presented in Fig. 3.

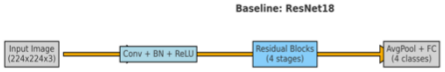

Fig. 3. ResNet Architecture

3.2.2 Proposed models of ResNet with variations

To enhance feature discrimination, we add CBAM in ResNet after each residual block in Fig. 4.

Fig. 4. ResNet Architecture

A comparison of the results obtained by the proposed models is presented in Section 5.

IV. EXPERIMENTS

4.1 Dataset

4.1.1 Source

Our dataset comes from Zou et al. (2022) introduced A100, the first systematic framework to evaluate the aesthetic ability of fashion compatibility models, including LAT (Liberalism Aesthetic Test) and AAT (Academicism Aesthetic Test). Because the task in the literature paper is the Fashion Matching problem. This is his dataset structure in Fig. 5.


```
[{
    "quesion_num": 1,
    "question": ["Pants_P00462138", "Shoes_P00447042", "Outwear_P00462123", "Bags_P00425745"],
    "answers": ["Top_P00440101", "Top_P00440104", "Top_P00278730", "Top_P00437254", "Top_P00440102"],
    "question_count": 4,
    "gt": 3,
    "gt_distribution": [0.054, 0.034, 0.832, 0.026, 0.054]
```


Fig. 5. Dataset Structure

It is not suitable for our image classification task. So, we manually annotated the data. The following label data is formed:

Image_Name	Label	Attribute
20250412_LAT_Clothing_1.jpg	1	Clothing
20250412_LAT_Bags_236.jpg	3	Bags
20250412_LAT_Shoes_100.jpg	0	Shoes
20250412_LAT_Accessories_30.jpg	2	Accessories

Fig. 6. Dataset Structure

4.1.2 Data analysis

The AAT dataset has 7,428 images with labels, and it suffers a severe class imbalance problem: Clothing occupies 75.7%, Shoes occupies 9.3%, Bags occupies 9.5%, and Accessories occupies just 5.6%. It is quite simple for the model to favor large categories and ignore small categories under this kind of imbalance.

4.1.3 Data processing

In order to solve the problem of class imbalance, some data processing was done. The processing includes resizing the image to 244×244, normalization, gray-scale conversion, and weight sampling using Weighted Random Sampler, improving the model's recognition ability for small categories.

4.2. The reason we chose epoch 19

As shown in Fig. 7, the training loss continues to decrease, while the validation loss drops quickly and kept around epoch 19. After this point, the validation curve becomes flat, with no further improvement, which shows the model has already learned the useful features. Although training accuracy remains high, further training beyond epoch 19 may lead to overfitting. To avoid this, we applied early stopping and selected epoch 19 as the optimal point. This model checkpoint was used as our baseline for comparison with improved versions such as AlexNet + CBAM, ensuring a fair and consistent evaluation.

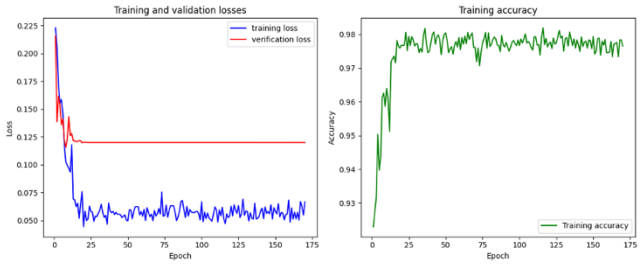

Fig. 7. Training and Validation loss

4.3. Evaluation strategy

We split the dataset into a training set of 7,427 images and a test set of 987 images. For model performance evaluation, we used standard classification metrics like Precision, Recall, F1-score, and Accuracy. Given the extreme class imbalance in the dataset, we used Recall and F1-score predominantly to give a better indication of the model's ability to recognize minority categories like Accessories and Bags.

4.4. Key model hyperparameters

In training, we set the learning rate to 0.0001 and applied a batch size of 32. The Adam optimizer was chosen because of the speed of convergence and stability.

These settings were used to achieve a balance between training efficiency and model generalization.

V. RESULTS

In this research, the performance of the nine proposed models and two baseline models was evaluated in our dataset. All methods were implemented on Colab. We buy the GPU service from Colab. Each model was run with 19 epochs.

5.1 The results on the AlexNet (change the number of convolutional layers or classifiers)

In this study, the effects of different numbers of convolutional layers and classifier configurations on the classification performance of fashion products were systematically compared. According to the experimental results, the 6Conv_SVM model (group7) achieved the best performance in all evaluation indicators, including Recall 0.8513, Precision 0.8722, F1 Score 0.8466, and Specificity 0.9393.

Compared with the baseline model AlexNet, 6Conv_SVM achieved significant improvements in classification accuracy and robustness. Further analysis shows that under the same number of convolutional layers, the model using the SVM classifier is overall better than the corresponding Softmax classifier model, especially in terms of recall and F1 score. In addition, appropriately increasing the number of convolutional layers can also help the model extract richer feature information, thereby improving the overall classification performance. In addition, when training efficiency is prioritized in real-world applications, 3Conv_SVM provides an excellent balance between performance and computational cost.

Table 3

Experimental Result(AlexNet)

Group	Recall	Precision	F1 Score	Specificity
1	0.8222	0.8525	0.8194	0.9309
2	0.8331	0.8479	0.8238	0.9322
3	0.8402	0.858	0.8384	0.9366
4	0.8343	0.8546	0.8274	0.9327
5	0.8214	0.857	0.8235	0.9301
6	0.8295	0.8519	0.82	0.9308
7	0.8513	0.8722	0.8466	0.9393

However, 6Conv_SVM (Fig. 8) shows a balance between precision and recall in the clothing and accessories categories, which indicates that there is still potential for further improvement in handling fine-grained categories.

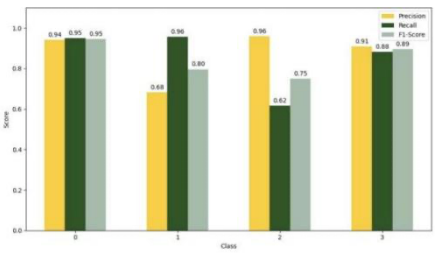

Fig. 8. Classification Report for 6Conv SVM

The confusion matrices (Fig. 9) of the seven experimental groups are shown, which shows that increasing the number of layers can improve the overall accuracy. Due to the diverse forms of accessories, the traditional improved Alexnet performs poorly in accessory classification. The following study 2 continues to solve this problem.

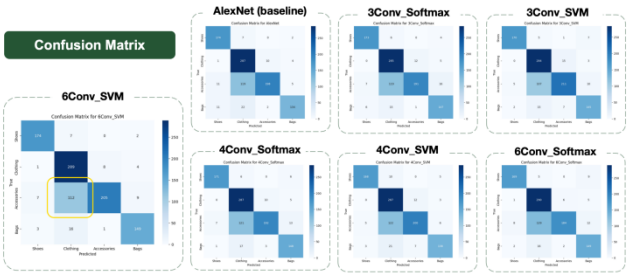

Fig. 9. Confusion matrices of AlexNet

5.2 The results on the AlexNet+CBAM

The proposed methods were tested on our dataset under specified experimental conditions. The results are listed in Table 4. Compared to the baseline AlexNet, which achieved an F1 score of 0.8194, the best-performing variant AlexNet_CBAM(16&7) reaches an F1 score of 0.8720.

Table 4

Experimental Result(AlexNet + CBAM)

Model	Recall	Precision	F1 Score
AlexNet(baseline)	0.8222	0.8525	0.8194
AlexNet_CBAM(16&7)	0.8810	0.8680	0.8720
AlexNet_CBAM(8&7)	0.8380	0.8350	0.8360
AlexNet_CBAM(8&5)	0.8420	0.8370	0.8380

In the baseline, “accessories” were often misclassified as “clothing,” but the CBAM-enhanced version reduced such misclassifications significantly. Specifically, AlexNet+CBAM (16&7) increases the correct predictions for “accessories” from 198 to 291.

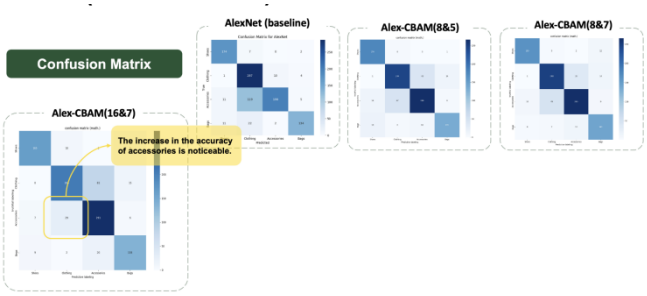

Fig. 10. Confusion matrices of AlexNet+CBAM

5.3 The results on the ResNet+CBAM

The results are listed in Table 5 & 6. The CBAM-enhanced model outperformed the baseline in every key metric. The overall accuracy improved from 85% to 93%, and the macro F1-score rose from 0.85 to 0.93. Most notably, the F1-score for Accessories increased from 0.79 to 0.91, showing that CBAM helped the model better distinguish smaller or visually similar categories.

Table 5

ResNet Result Report.

	Precision	Recall	F1 Score
Shoes	0.84	0.97	0.9
Clothing	0.88	0.94	0.91
Accessories	0.93	0.68	0.79
Bags	0.72	0.89	0.8
Accuracy			0.85
Marco avg	0.84	0.87	0.85
Weighted avg	0.86	0.85	0.85

Table 6

ResNet+CBAM Result Report.

	Precision	Recall	F1 Score
Shoes	0.95	0.99	0.97
Clothing	0.91	0.96	0.93
Accessories	0.94	0.87	0.91
Bags	0.91	0.89	0.9
Accuracy			0.93
Marco avg	0.93	0.93	0.93
Weighted avg	0.93	0.93	0.92

5.4 Summary

Table 7 compares all the models we trained in this study. The result shows that simply increasing or reducing the number of layers does not lead to a significant performance boost. Attention modules like CBAM do improve model performance, especially for complex or minority classes. And when combined with deeper networks like ResNet, the improvement becomes even more significant.

Table 7

Experimental Result summary

Model	Recall	Precision	F1 Score
AlexNet (baseline)	0.8222	0.8525	0.8194
3Conv_Softmax	0.8331	0.8479	0.8238
3Conv_SVM	0.8402	0.858	0.8384
4Conv_Softmax	0.8343	0.8546	0.8274
4Conv_SVM	0.8214	0.857	0.8235
6Conv_Softmax	0.8295	0.8519	0.82
6Conv_SVM	0.8513	0.8722	0.8466
Alex-CBAM(16&7)	0.881	0.868	0.872
Alex-CBAM(8&7)	0.838	0.835	0.836
Alex-CBAM(8&5)	0.842	0.837	0.838
Resnet (baseline)	0.84	0.87	0.85
Resnet-CBAM	0.93	0.93	0.93

VI. CONCLUSION

In this project, we implemented and compared three approaches to improve fashion item classification: scaling AlexNet, introducing CBAM, and combining CBAM with ResNet18.

In this project, we experimented with different methods to increase fashion item classification, including scaling AlexNet, incorporating the CBAM attention module, and combining CBAM with ResNet18. Our main contribution is demonstrating that attention mechanisms, particularly when used in conjunction with deeper networks, can significantly boost classification accuracy for small and complex categories.

The most important strength of our solution is its strong overall performance and better generalization, especially in the detection of minority classes like bags and accessories. The system using the implementation of CBAM and ResNet18 achieved 93% accuracy and F1-scores of more than 0.90 for all categories.

However, our current work also has its limitations. CBAM sometimes over-emphasizes local textures, resulting in occasional misclassification. Additionally, the deeper architectures used by us require more computational resources, which may not be ideal for real-time applications.

With further time, future improvements can include exploring CNN-Transformer hybrid models to extract global features more effectively, and optimizing attention modules with lighter alternatives like SE or ECA to reduce computational complexity.

REFERENCES

[1] Chaudhuri A, Messina P, Kokkula S, et al. A smart system for selection of optimal product images in e-commerce[C]//2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018: 1728-1736.

[2] George M, Floerkemeier C. Recognizing products: A per-exemplar multi-label image classification approach[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13. Springer International Publishing, 2014: 440-455.

[3] Hakami N A. Unveiling Online Shopping Quality: Taxonomy-Based Multi-label Classification of Shopping App Reviews[C]//IBIMA Conference on Artificial intelligence and Machine Learning. Cham: Springer Nature Switzerland, 2024: 157-173.

[4] Abbas W, Zhang Z, Asim M, et al. Ai-driven precision clothing classification: Revolutionizing online fashion retailing with hybrid two-objective learning[J]. Information, 2024, 15(4): 196.

[5] Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1096-1104).

[6] Corbiere, C., Ben-Younes, H., Ramé, A., & Ollion, C. (2017). Leveraging weakly annotated data for fashion image retrieval and label prediction. In Proceedings of the IEEE international conference on computer vision workshops (pp. 2268-2274).

[7] Zou, X., Pang, K., Zhang, W., & Wong, W. (2022). How good is aesthetic ability of a fashion model?. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 21200-21209).

[8] Dai Y, Tao J, Ouyang C, et al. Clothing recognition based on improved resnet18 model[C]//2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). IEEE, 2022, 5: 1297-1301.

[9] Xuan X, Han R, Ji S, et al. Research on Clothing Image Classification Models Based on CNN and Transfer Learning[C]//2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2021, 5: 1461-1466.

[10] Jha B K. E-commerce product image classification using transfer learning[C]//2021 5th International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2021: 904-912.

[11] K. Shah and K. Deulkar, "Lightweight Apparel Classification with Residual and Inverted Residual Block based Architectures," 2021 IEEE Cloud Summit (Cloud Summit), Hempstead, NY, USA, 2021, pp. 57-62, doi: 10.1109/IEEECloudSummit52029.2021.00017.[13]

[12] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[13] Eldem H, Ülker E, Işıklı O Y. Alexnet architecture variations with transfer learning for classification of wound images[J]. Engineering Science and Technology, an International Journal, 2023, 45: 101490.

[14] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. *European Conference on Computer Vision (ECCV)*.

[15] Samir, S., Emary, E., El-Sayed, K., & Onsi, H. (2020). Optimization of a pre-trained AlexNet model for detecting and localizing image forgeries. Information, 11(5), 275.