applied sciences

MDPI

Article

A Study on Deep Learning Performances of Identifying Images' Emotion: Comparing Performances of Three Algorithms to Analyze Fashion Items

Gaeun Lee [1], Seoyun Yi [2] and Jongtae Lee [1,*]

[1] Department of Business Administration, Seoul Women's University, Seoul 03079, Republic of Korea; lgy3093@swu.ac.kr
[2] Department of Data Science, Seoul Women's University, Seoul 03079, Republic of Korea; winterlike13@swu.ac.kr
* Correspondence: light4u@swu.ac.kr

Abstract: Emotion recognition using AI has garnered significant attention in recent years, particularly in areas such as fashion, where understanding consumer sentiment can drive more personalized and effective marketing strategies. This study aims to propose an AI model that automatically analyzes the emotional emotions of fashion images and compares the performance of CNN, ViT, and ResNet to determine the most suitable model. The experimental results showed that the vision transformer (ViT) model outperformed both ResNet50 and CNN models. This is due to the fact that transformer-based models, like ViT, offer greater scalability compared to CNN-based models. Specifically, ViT utilizes the transformer structure directly, which requires fewer computational resources during transfer learning compared to CNNs. This study illustrates that vision transformer (ViT) demonstrates higher performances with fewer computational resources than CNN during transfer learning. For academic and practical implications, the strong performance of ViT demonstrates the scalability and efficiency of transformer structures, indicating the need for further research applying transformer-based models to diverse datasets and environments.

Keywords: vision transformer; CNN; ResNet; emotion forecast; artificial intelligence

check for updates

Academic Editor: Pedro Couto

Received: 31 December 2024
Revised: 23 February 2025
Accepted: 10 March 2025
Published: 18 March 2025

Citation: Lee, G.; Yi, S.; Lee, J. A Study on Deep Learning Performances of Identifying Images' Emotion: Comparing Performances of Three Algorithms to Analyze Fashion Items. *Appl. Sci.* **2025**, *15*, 3318. https://doi.org/10.3390/app15063318

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Research Background and Topic

Emotion recognition using AI has garnered significant attention in recent years, particularly in areas such as fashion, where understanding consumer sentiment can drive more personalized and effective marketing strategies. The demand for emotional analysis and image-based recommendation systems in the fashion industry has been rapidly growing. Nonetheless, the emotion analysis was accepted in the studies on consumers' emotions about fashion brands traditionally [1,2]. Small fashion brands, in particular, face difficulties in utilizing digital content effectively due to limited resources. To address these challenges, AI-based technologies have emerged as a solution. Analyzing and classifying fashion images, especially in terms of their emotional emotions, is a highly complex task that heavily relies on the performance of advanced deep learning models. Chakriswaran et al. suggest that sentiment analysis refers to predicting the emotion and attitude of an individual person underlined in the context and content. Commonly, the sentiment analysis research tends to focus on not only the text contents but also the emotional scores of images [3].

Previous studies on image analysis have been adopted successfully. Convolutional neural networks (CNNs) and newer studies also introduce that vision transformers (ViT) and residual networks (ResNet) are widely used architectures for image analysis. CNNs have proven effective in learning features from fashion images and performing well in classification tasks. ViTs, on the other hand, excel in processing broader image information efficiently, utilizing resources optimally. ResNet, developed to solve the vanishing gradient problem in deeper networks, performs exceptionally well in analyzing complex images due to its deep learning capabilities.

This study aims to propose an AI model that automatically analyzes the emotional emotions of fashion images (e.g., sexy, classic, calm) and compares the performance of CNN, ViT, and ResNet to determine the most suitable model. This study will assist fashion companies in improving marketing strategies through emotional analysis and providing content tailored to consumer preferences.

1.2. Related Works

CNN has been used for image training, image recognition, and lots of image-related tasks. In Krizhevsky et al.'s paper, the AlexNet model was created based on CNN, and it is famous for increasing the accuracy of ImageNet Top-5 [4]. Also, Tabibu et al. conducted a study about classifying subtypes of pan-renal cell carcinoma (CNN) using CNN. In this study, the model based on CNN could successfully distinguish RCC from normal tissues, and the accuracy of the model exceeded 90% with excellent results [5]. In this way, CNN can be a tool that shows high performance for image-related tasks, and it is thought to be suitable for the fashion image-related and emotion-related tasks conducted in this study.

ViT also has been widely utilized to perform quantities of image-related tasks. Abd Alaziz et al. made the fashion system, which reinforces fashion classification using ViT [6]. In this reference, it says that ViT is a set of transformer blocks that consists of multi-layer perceptron layers and multi-head attention layers [6]. So, with this characteristic, the fashion system made great results, such as showing over 90% accuracy in various factors. Furthermore, Abd Alaziz did a comparative study between CNN-based models and the ViT-based model in this reference, and the latter one showed better performances [6]. Therefore, these points suggest that ViT can be a good ingredient for fashion emotion analysis conducted in this paper.

ResNet has proven especially successful in tasks such as image classification and object detection, which are crucial in analyzing fashion images. He et al. demonstrated the strength of ResNet in the ImageNet competition, where ResNet-152 outperformed shallower networks, achieving state-of-the-art performance in image classification [7]. The ability to maintain accuracy in deep networks made ResNet one of the most widely adopted architectures for visual recognition tasks [7]. The research suggested that deep learning models, like ResNet, could provide significant improvements in complex image analysis tasks, such as those in the fashion industry.

1.3. Structure and Research Process

This study addresses the topic in the following order: "Algorithms and Neural Network Definitions", which introduces each definition of the algorithms chosen for this experiment, "Materials and Methods", which suggests the comparison process of image analysis algorithms, "Experimental Results and Interpretation", to explain the detailed research analysis results, and "Conclusions". To accomplish the research purpose, this study suggests the following research model (Figure 1).



Figure 1. Research process of this study.

2. Algorithms and Neural Network Definitions

Diverse studies on emotion recognition in images typically involve analyzing visual features such as facial expressions, color schemes, and clothing styles and have employed deep learning models to analyze visual elements including colors, design patterns, and styles to decide emotional features. Previous studies, including Verma and Verma, suggest that brightness of colors tends to affect the perceived emotions and clothing styles, and accessories can affect the emotional feelings for the customers [8]. These studies introduce how emotion recognition can be possible with AI models categorizing emotions in fashion images for commercial purposes and how emotional factors in fashion item images can be optimized to suggest stronger emotional features to the customers [8].

2.1. CNN (Convolutional Neural Network)

Convolutional neural network (CNN) is a model that utilizes layers with convolving filters applied to local features, and it was first introduced by Lecun et al. in the form of "LeNet". After that, deep learning, including CNN, became a hot topic because of CNN-based "AlexNet" proposed by A. Krizhevsky et al. [4,9].

CNN is the neural network usually used for image training, image recognition, and video recognition. CNN has great feature compatibility, and it could identify highly relevant features automatically without supervision [10]. Also, the network could improve generalization and prevent overfitting [10]. Nonetheless, as CNN is becoming more advanced nowadays, it is expected that resource consumption will be high because large amounts of data and massive computing power are needed in deeper CNN [11,12] (Figure 2).

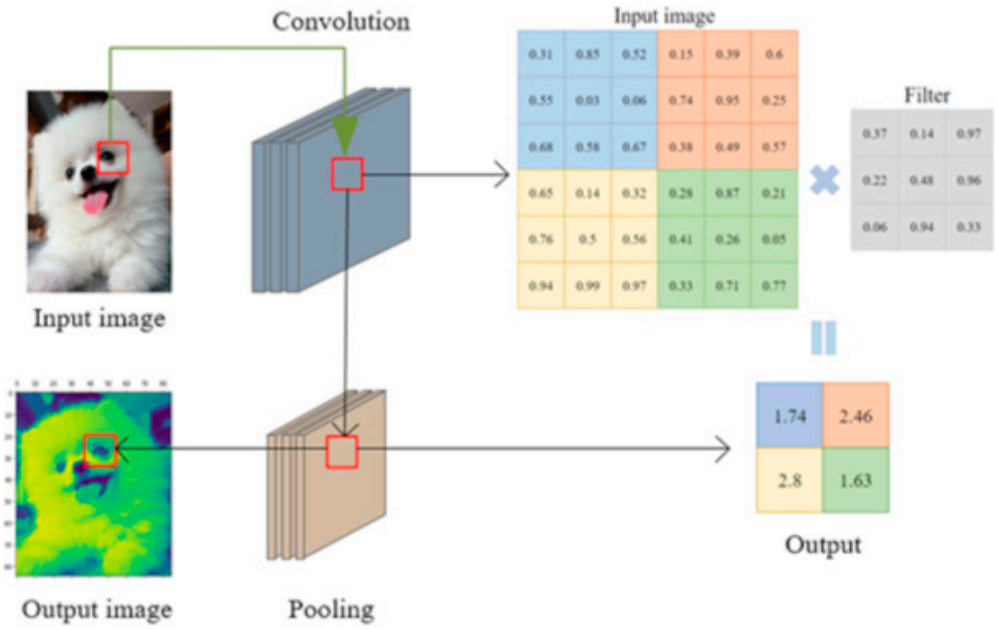

Figure 2. Process of CNN algorithm to analyze Images [11].

2.2. ViT (Vision Transformer)

Transformer has been widely used for NLP research, and it is structured with high computational efficiency and scalability. Nonetheless, it was not applied well to computer

vision tasks [13]. Hence, for being well applied to computer vision tasks, visual transformer (ViT) was made, which is a model composed of a standard transformer and the image itself directly inserted in it [14] (Figure 3).

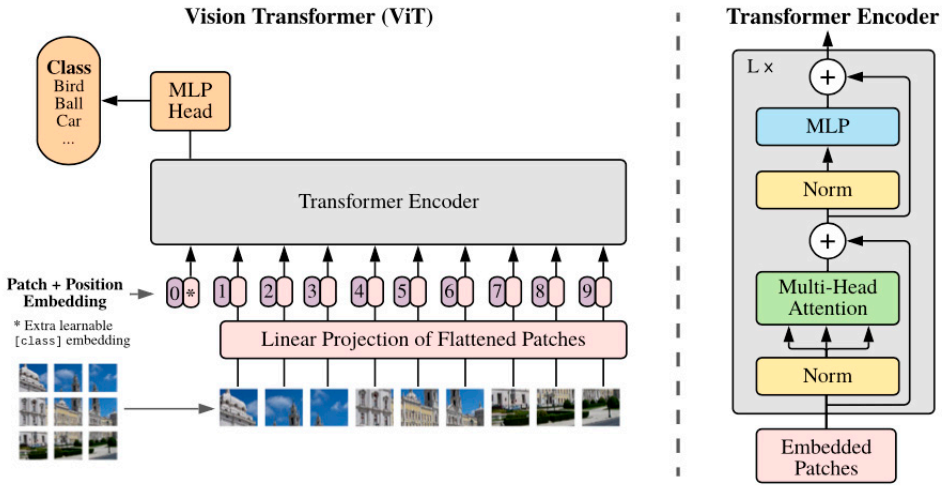

Figure 3. Process of vision transformer algorithm to analyze images [14].

ViT has great scalability because it uses the transformer structure almost as is, and it uses much fewer computational resources for training than CNN in TL (transfer learning). Also, as transformer has proven that it has superior performance in large-scale learning, we could think that ViT could show great performance [13,15] (Figure 4).

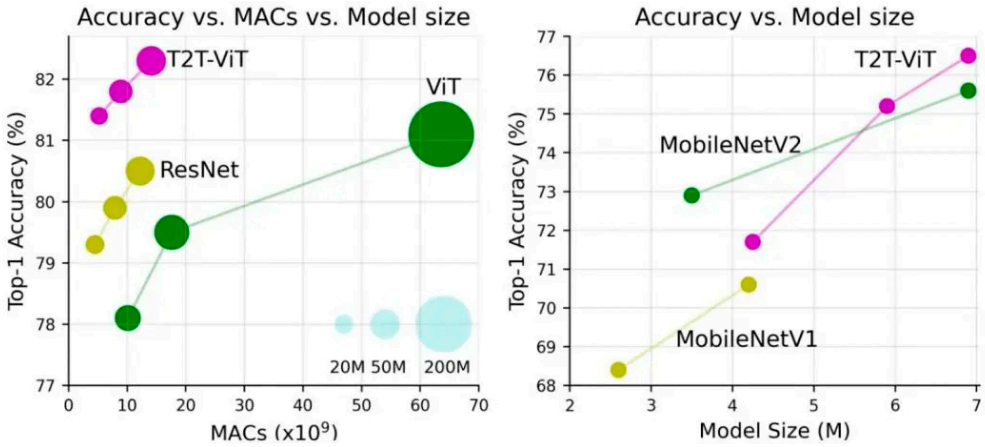

Figure 4. Comparing the performances of image analysis algorithms [15].

Nonetheless, ViT requires more data than CNN because of the lack of inductive bias, which is an additional assumption used to predict output for unseen input. Moreover, if ViT is learned with an insufficient amount of data, generalization performance will decrease [13].

2.3. ResNet (Residual Network)

ResNet (residual network) is a deep learning model proposed by Kaiming He et al. designed to maintain training efficiency even with a significant increase in the depth of the neural network [7]. The core idea of ResNet is to address the vanishing gradient problem through "residual learning" [7]. While traditional networks become increasingly difficult to train as their layers deepen, ResNet is designed to learn the difference between the input and output values at each layer, known as the residual, enabling effective training even in deeper networks [7].

ResNet is widely used in computer vision tasks such as image classification, object detection, and segmentation. Its outstanding performance, particularly demonstrated on the ImageNet dataset, has led to its adoption as a state-of-the-art model across various applications. Additionally, the residual block structure of ResNet has been applied to other deep learning architectures, extending its use beyond vision to natural language processing and speech recognition fields.

A major advantage of ResNet is its ability to learn without performance degradation while expanding the depth of the network. This allows for the construction of deeper and more complex models compared to traditional CNN architectures, providing exceptionally high performance in image recognition and classification. It is also conducive to transfer learning and generally converges faster than other models. Nonetheless, ResNet has the drawback of consuming substantial computational resources when the model's depth becomes very high. Additionally, the basic ResNet model may exhibit limited performance on more complex datasets or highly variable images, prompting the proposal of various extended models to address these limitations. The ResNet algorithm can demonstrate higher forecast and analysis performance with the adoption of a residual block, while the previous neural network models, including CNN, demonstrate lower performance when the network depths become deeper [16] (Figure 5).

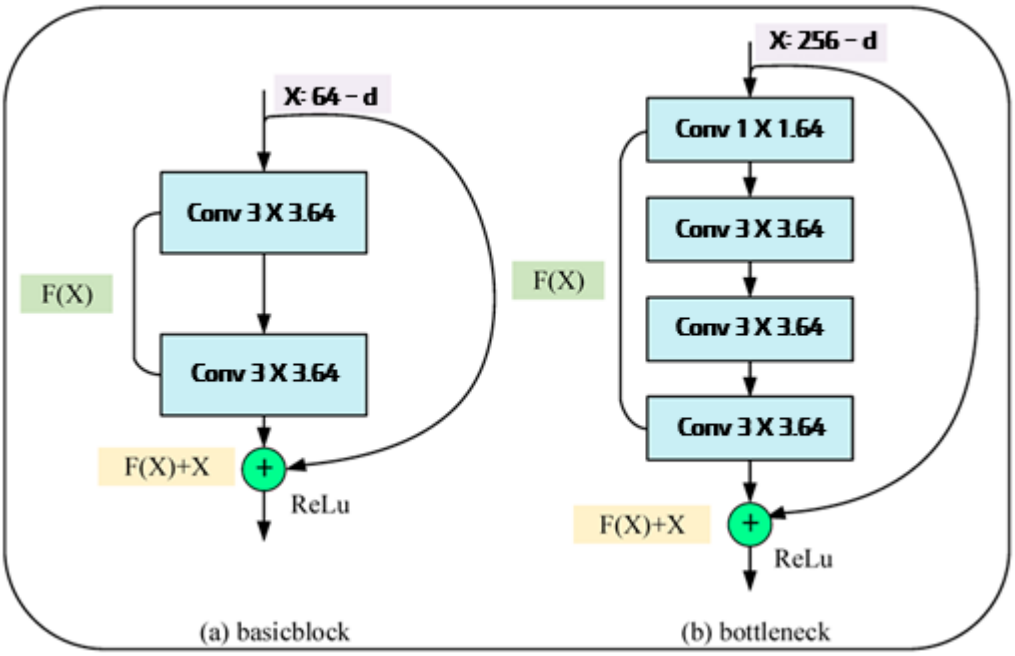

Figure 5. Residual block structure of ResNet algorithms (ResNet-18 and ResNet-34) [16].

2.4. CoAtNet (Convolution and Self-Attention Network)

As mentioned earlier, CNN, ResNet, and ViT have their specific advantages and disadvantages far from each other. For instance, ViT can support an advantage in terms of data size compared to the CNN algorithm. However, ViT can be less effective than CNN without the support of well-designed large-scale dataset in the training stage. According to these considerable features, there is diverse research to overcome those disadvantages, such as CoAtNet [17,18]. CoAtNet and ResNet-Vit are newly suggested algorithms to converge the technological concepts of ViT and CNN or ViT and ResNeet to generalize the inductive bias. This newer research can suggest higher performances to analyze images, but still it is not easy to expect the practical performances to be permanent. Specifically, according to Dai et al. [17], flops and params can be core factors to generate higher performance. In the research of Dai et al. [17], the accuracy of CvT shows lower accuracy performance until

the params were under 300 million, but the slope was higher than the slope of CoAtNet, and it can be expected that the performance of CvT can exceed that of CoAtNet in a certain circumstance.

3. Materials and Methods

3.1. Dataset Description and Preprocessing

The dataset used in this study comprises two types: an image dataset and an emotion score dataset. First, the image dataset was provided by the "A" company, an SME fashion company for women's apparel and fashion items in South Korea. The "A" company tries to analyze its fashion items without MD's handling or minimizing the MD's personal favors. The research team (the authors) adopts the algorithms of ViT, CNN, and ResNet to analyze the emotional features of the fashion images after the training process. The dataset consists of 1169 fashion images (totally 1.09 GB) categorized into four categories such as bottom (pants, skirt), dress, outer (cardigan, coat, jacket), and top (blouse, knit, shirt, T-shirt, vest). These images primarily depict full-body shots of female models with studio or street backgrounds. In this study, the backgrounds of each image were removed to extract only the full-body models. The rembg library in Python was used to make the background transparent. The file extension of the images was then changed from JPG to PNG to reflect the transparent backgrounds (Figure 6).



Figure 6. (**a**) Is an original image and (**b**) is an image with transparent background.

Next, the emotion score dataset was assigned by a fashion MD (merchandiser) at the "A" Company(initialized) who rated the emotion of each image (<Table 1>). The emotion scores encompass a total of 22 elements across 11 pairs: <Decorated, Classic>, <Fancy, Simple>, <Sexy, Elegant>, <Lively, Demure>, <Energetic, Calm>, <Light, Dark>, <Soft, Hard>, <Practical, Premium>, <Party Look, Daily Life Look>, <Stand-out, Ordinary>, <Feminine Mannish> (<Table 2>). Each pair's scores were assigned from 0 to 10, with the total score for each pair summing up to 10. For instance, if the original score of the factor "Well-Decorated" is 6, the original score of the factor "Classic" is 4. After analyzing all the data, to suggest the recommended emotion, estimated emotion scores of 5 or higher were

converted to 1—this means that the image suggests the emotion, while scores below 5 were set to 0 for the model training process. For instance, if the original score of the factor "Soft" exceeds 5, the dummy variable for the factor would be marked as one. All the descriptions of the factors are illustrated in <Table 3>.

Table 1. Brief information of dataset.

First Category	Second Category	Amount of Goods
Bottom	Pants	94
	Skirt	134
Outer	Cardigan	120
	Coat	120
	Jacket	120
Top	Blouse	112
	Knit	118
	Shirt	109
	Vest	58
	T-Shirt	104
Dress	Dress	80
Total		1169

Table 2. Emotion pairs for the study.

Emotional Pairwise Comparison		Paring Scores (If a Is 10, a's Pair = 0)	Dummy Variables (1: Pairing Score > 5, 0: Paring Score < 5)
Well-Decorated	Classic	Totally 10	0/1
Fancy	Simple	Totally 10	0/1
Sexy	Elegant	Totally 10	0/1
Lively	Demure	Totally 10	0/1
Energetic	Calm	Totally 10	0/1
Light	Dark	Totally 10	0/1
Soft	Hard	Totally 10	0/1
Practical	Premium	Totally 10	0/1
Party Look	Daily Life Look	Totally 10	0/1
Stand-out	Ordinary	Totally 10	0/1
Feminine	Mannish	Totally 10	0/1

Table 3. General experiment parameter definition for the study.

Features	Definition for the Study
EPOCH	20 times for each epoch X 20 epochs
Data Set	Totally 1169 Fashion Wear Images with Four Categories of Top/Bottom/Outer/Dress
Machine Learning Environment	Python 3.10 and PyTorch 2.10 at Google Colaboratory Lab with GPU
Optimizer and Activation Function	ADAM/Sigmoid
Training: Testing Ratio	Training 80: Testing 20

3.2. Model Design and Settings

For this study, three algorithms are adapted for making the models: CNN, ViT, and ResNet50. Since CoAtNet and ResNet-ViT are closer to converged versions of the three core algorithms to be compared, they are excluded from this study. In order to study and make models under the same conditions, we set the six factors equally. First, we each set "Steps per epoch" and "Epoch" for 20 times, and we used the "Adam" optimizer for modeling. Also, "Activation" was set to "Sigmoid", "Loss" was set to "Binary Cross Entropy", and "Metrics" was set to "Accuracy". The independent variable for training the model was designated as fashion image data from the company, and we designated the dependent variable as the dataset containing 11 pairs of emotions presented by the company. The dataset was divided into a 7:3 ratio, 814 for train data and 355 for test data. After training the model with these variables, when a random image was inputted into the model, the model printed the score (between 0 and 10) out for each emotion for the image.

4. Experimental Results and Interpretation

This study compares the results from three models, each based on CNN, ViT, and ResNet50. There were no significant differences in the quantitative performance parts, such as the accuracy of those models, because of the qualitative characteristics of this study. Therefore, this study adopts the performances to be expressed through loss values in this study because the expectation performances of continuous variables are not easily measured compared to the expectation performances of discrete variables, so we focused on comparing the loss rate values in each model. The loss rate variable can illustrate the expectation/forecast errors of the training data and the testing data for each algorithm, and previous studies, including the study of Pham et al., suggest the lowest loss rate can be considered after repeating the epoch on the model to illustrate the forecast performances of the algorithms [19].

$$\mathrm{W} = \mathrm{Error}(f_\theta(x),\ y) \tag{1}$$

$$\mathrm{W} = \text{Loss Function},\ \theta = \text{factor weights} \tag{2}$$

In the analysis result, the minimum loss value each epoch (min_loss) of CNN was 0.3987, and the final validation loss of it was 0.4157. In the case of ViT, min_loss was 0.0038, and the final validation loss was 0.1438. ResNet50 got 0.3225 for min_loss value and 0.718 for the final validation loss. ViT got the lowest min_loss value and the final validation loss value among the models. Min_loss of CNN was similar to ResNet50's, but the final validation loss of CNN was much lower than ResNet50's (<Table 4>). Figure 7 illustrates the performances comparison of each model.

Table 4. Loss rates of each model.

Algorithm	Min Loss per Each Epoch	Validation Loss
CNN	0.3987	0.4157
ViT	0.0038	0.1438
ResNet50	0.3225	0.718

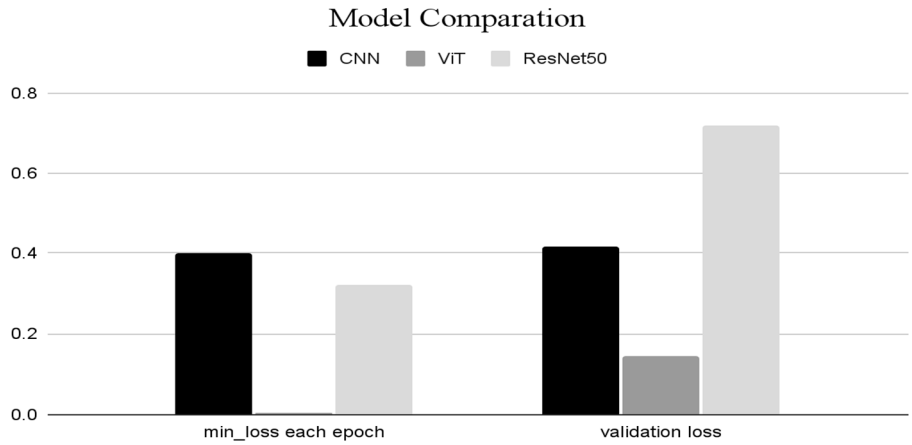

Figure 7. Performances comparison of each model.

After comparing the algorithms and selecting the highest performance algorithm (ViT), this study demonstrates the comparisons of AI's emotion forecasts and MD's emotion scores. According to the comparison results, the emotion suggested by the algorithm demonstrates more than 80% similarity, which means that 8 to 11 pairs of emotions forecasted by the AI algorithm would be the same as the MD's emotional recommendation (Figure 8).



Figure 8. Recommended emotion comparison of AI and human MD.

5. Conclusions

5.1. Summary of Experimental Results and Implications

The experimental results showed that the vision transformer (ViT) model outperformed both ResNet50 and CNN models. This is due to the fact that transformer-based models, like ViT, offer greater scalability compared to CNN-based models. This result supports previous studies comparing the analysis performance of ViT and CNN variants [20,21]. In this study, ViT demonstrates a relatively lower loss rate, which highlights its ability to analyze images with complex fashion items effectively. Specifically, ViT directly utilizes the transformer structure, which requires fewer computational resources during transfer learning compared to CNN variants—such as ResNet and CNN in this study. On the other hand, CNN-based models tend to require larger datasets due to their increasingly deep architectures. This study does not consider the analysis speeds of each model, as these can vary depending on experimental conditions, including virtual GPU performance, network capabilities, and other factors. The dataset used in this study may not have been large enough to fully train these models, suggesting that dataset size plays a significant role in model performance.

For academic implications, first, the correlation between dataset size and model performance should be stressed. This study revealed that dataset size can be considered as a meaningful factor in determining model performances and choosing more proper models. This suggests that in future research, CNN-based models may demonstrate improved performance when larger datasets are used. This study also supports the study of Mauricio et al. to illustrate the higher performance of ViT for image classification than traditional CNN [22].

Second, the scalability of transformer architectures. The strong performance of ViT demonstrates the scalability and efficiency of transformer structures, indicating the need for further research applying transformer-based models to diverse datasets and environments. Transformer architectures have the potential to be applied to various data types, highlighting their potential for broader use in future studies.

Third, this study suggests the academic consideration to consider the capability of ViT, which still is a newer algorithm. This study can be academic evidence of the superior performance of ViT.

For practical implications, first, the efficiency of transformer-based models can be considered for the business performances. Vision transformer (ViT) uses fewer computational resources than CNN during transfer learning, making it more efficient for practical applications [22]. In particular, ViT demonstrates strong performance even when using limited datasets, making it an effective solution in industrial settings where data resources are scarce. With the ViT algorithm, small and medium-sized corporations can be supported to enhance their business performances, especially business automation and recommendations for the customers.

Second, the importance of dataset size. CNN-based models, due to their increasingly deep architectures, require large datasets. Therefore, when applying these models in practice, careful consideration must be given to the relationship between dataset size and model learning efficiency. Insufficient datasets can hinder the performance of CNN models.

Third, SMEs (small and medium-sized enterprises), including the fashion industry, should face the comparatively technological handicaps of big companies because of the lack of manpower, investment abilities, understanding levels on convergence technologies, and others. According to the research results, ViT and other image processing algorithms may help the handicapped SMEs to save time, salaries, and investment portfolios. For instance, with ViT as this study, fashion SMEs can reduce the spending time to realign and combine the fashion items.

5.2. Limitations of the Sutdy and Future Research Suggestions

This study compares the performance of CNN, ViT, and ResNet models in analyzing the emotional emotions of fashion images. Nonetheless, several limitations exist.

First, the diversity of the dataset is somewhat limited, potentially skewing the results toward specific fashion styles or emotions. Future research should incorporate a larger and more diverse dataset, including varied fashion images and backgrounds. Also, though the dataset consists of practical fashion items in an online shopping service, a relatively smaller dataset may be relatively smaller to limit the generalizability of the results. For the future studies, it can be considered to adopt larger datasets such as ImageNet. Also, the future studies should consider adopting the algorithms and larger datasets to understand and generalize the user's responses—for instance, the impacts of generations, genders, and personal fashion favors.

Second, although the performance of the models was generally strong, emerging algorithms such as CoAtNet, which combine the local feature extraction capabilities of CNNs and the global information processing strengths of ViTs, have not been sufficiently explored. CoAtNet is a hybrid model that merges the strengths of CNN and ViT, offering enhanced performance in various image analysis tasks. Incorporating CoAtNet in future studies could provide further insights into its effectiveness in fashion image emotion analysis, broadening the scope beyond CNN, ViT, and ResNet.

Third, this study adopts the analysis environment of Google Colaboratory Lab with the GPU enable option. Nonetheless, in practical business situations, there can be different and diverse system environments affecting the performances. Also, the diversities of the fashion items, such as categories, MD's personal scoring results, and the complexities of the images should be considered.

Fourth, there are diverse performance indices to compare the results of machine learning algorithms. This study mainly considers the loss function results as standing out the performances of each algorithm following the study of Muthukumar et al. because this study converted the emotional scores into 0–1 selection [23]. In future studies, diverse indices can be considered, though.

Lastly, while this study evaluated model performance using standard metrics such as accuracy, precision, and recall, it did not fully account for subjective user experiences or emotional responses. Future studies should integrate user feedback and subjective evaluation metrics to assess the practical applicability and emotional impact of these models more comprehensively.

Advancing research in this direction could lead to more refined and practical AI models for fashion image analysis.

Author Contributions: Writing—original draft, G.L. and S.Y.; Writing—review & editing, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a research grant from Seoul Women's University (2025-0023).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the corresponding author's database upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Giri, C.; Harale, N.; Thomassey, S.; Zeng, X. Analysis of Consumer Emotions about Fashion Brands: An Exploratory Study. *Data Sci. Knowl. Eng. Sens. Decis. Support* **2018**, *11*, 1567–1574.

2. Yeo, S.F.; Tan, C.L.; Kumar, A.; Tan, K.H.J.; Wong, J.K. Investigating the Impact of AI-Powered Technologies on Instragrammers' Purchase Decisions in Digitalization Era—A Study of the Fashion and Apparel Industry. *Technol. Forecast. Soc. Chang.* **2022**, *177*, 121551. [CrossRef]
3. Chakriswaran, P.; Vincent, D.R.; Srinivasan, K.; Sharma, V.; Chang, C.-Y.; Reina, D.G. Emotion AI-Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues. *Appl. Sci.* **2019**, *9*, 5462. [CrossRef]
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2012**, *60*, 84–90. [CrossRef]
5. Tabibu, S.; Vinod, P.K.; Jawahar, C.V. Pan-Renal Cell Carcinoma Classification and Survival Prediction from Histopathology Images Using Deep Learning. *Sci. Rep.* **2019**, *9*, 10509. [CrossRef] [PubMed]
6. Abd Alaziz, H.M.; Elmannai, H.; Saleh, H.; Hadjouni, M.; Anter, A.M.; Koura, A.; Kayed, M. Enhancing Fashion Classification with Vision Transformer (ViT) and Developing Recommendation Fashion Systems Using DINOVA2. *Electronics* **2023**, *12*, 4263. [CrossRef]
7. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
8. Verma, G.; Verma, H. Hybrid-Deep Learning Model for Emotion Recognition Using Facial Expressions. *Rev. Socionetwork Strateg.* **2020**, *14*, 171–180. [CrossRef]
9. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
10. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *J. Big Data* **2021**, *8*, 53. [CrossRef]
11. Zhao, L.; Zhang, Z. A Improved Pooling Method for Convolutional Neural Networks. *Sci. Rep.* **2024**, *14*, 1589. [CrossRef]
12. Zhao, S.; Duan, Y.; Zhang, B. A Deep Learning Methodology Based on Adoptive Multiscale CNN and Enhanced Highway LSTM for Industrial Process Fault Diagnosis. *Reliab. Eng. Syst. Saf.* **2024**, *249*, 110208. [CrossRef]
13. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey of Vision Transformers. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
14. Thisanke, H.; Deshan, C.; Chamith, K.; Seneviratne, S.; Vidanaarachchi, R.; Herath, D. Semantic Segmentation Using Vision Transformers: A Survey. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106669. [CrossRef]
15. Viso.AI. Vision Transformers (ViT) in Image Recognition—2024 Guide. Available online: https://viso.ai/deep-learning/vision-transformer-vit/ (accessed on 24 December 2024).
16. Li, X.; Xu, X.; He, X.; Wei, X.; Yang, H. Intelligent Crack Detection Method Based on GM-ResNet. *Sensors* **2023**, *23*, 8369. [CrossRef]
17. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. CoAtNet: Marrying Convolution and Attention for All Data Sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977. [CrossRef]
18. Asif, M.; Rajab, T.; Hussain, S.; Rashid, M.; Wasi, S.; Ahmed, A.; Kanwal, K. Performance Evaluation of Deep Learning Algorithm Using High-End Media Processing Board in Real-Time Environment. *J. Sens.* **2022**, *2022*, 6335118. [CrossRef]
19. Pham, T.C.; Luong, C.M.; Hoang, V.D.; Doucet, A. AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function. *Sci. Rep.* **2021**, *11*, 17485. [CrossRef]
20. Alayón, S.; Hernández, J.; Fumero, F.J.; Sigut, J.F.; Díaz-Alemán, T. Comparison of the Performance of Convolutional Neural Networks and Vision Transformer-Based Systems for Automated Glaucoma Detection with Eye Fundus Images. *Appl. Sci.* **2023**, *13*, 12722. [CrossRef]
21. Khan, A.; Rauf, Z.; Sohail, A.; Khan, A.R.; Asif, H.; Asif, A.; Farooq, U. A survey of the vision transformers and their CNN-transformer based variants. *Artif. Intell. Rev.* **2023**, *56*, 2917–2970. [CrossRef]
22. Maurício, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Appl. Sci.* **2023**, *13*, 5521. [CrossRef]
23. Muthukumar, V.; Naran, A.; Subramanian, V.B.; Belkin, M.; Hsu, D.; Sahai, A. Classification vs Regression in Overparameterized Regimes: Does the Loss Function matter? *J. Mach. Learn. Res.* **2021**, *22*, 1–69.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.