Enhanced Convolutional Neural Network for Fashion Classification

Lailan M. Haji
Computer Science Department, University of Zakho, Duhok, Iraq
lailan.haji@uoz.edu.krd

Omar M. Mustafa
Computer Science Department, University of Zakho, Duhok, Iraq
omar.mustafa@uoz.edu.krd

Sherwan A. Abdullah
College of Education, University of Zakho, Duhok, Iraq
sherwan.abdullah@uoz.edu.krd

Omar M. Ahmed
Computer Information System Department, Duhok Polytechnic University, Iraq
omar.alzakholi@dpu.edu.krd (corresponding author)

Received: 15 June 2024 | Revised: 8 July 2024 | Accepted: 21 July 2024

ABSTRACT

Fashion items are hard to classify since there are a million variations in style, texture, and pattern. Image classification is among the noted strengths of convolutional neural networks. This research introduces an improved CNN architecture for fashion classification, utilizing image augmentation and batch normalization to improve model performance and generalization. To make the model more robust, image augmentation techniques like rotation, width and height shift, zoom, and flips were employed. In addition, a Batch Normalization layer is added in the middle, which can help on stabilizing the learning process and accelerating convergence. The proposed model was trained on an augmented dataset, achieving a satisfactory improvement in test accuracy of 91.97% compared to a baseline CNN model, which obtained 88.5% accuracy. According to the results, the image augmentation with the application of Batch Normalization improves the CNN architecture for better effectiveness in fashion classification tasks.

Keywords-convolutional neural networks; batch normalization; image augmentation; fashion classification

I. INTRODUCTION

The classification of fashion is considered as an essential task for computer vision, which might cover applications in e-commerce, retail, inventory management, and trend analysis. It will help classify and identify fashion items from their respective images and thus improve user experience with more efficient operations and other valuable consumer insights. The problem of classifying fashion remains very challenging since clothing classes involve vast diversities in styles, patterns, and textures [1]. The Convolutional Neural Networks (CNNs) has recently been the most prevalent method for image classification and has given excellent results in many application areas. The CNN is designed to automatically and adaptively learn spatial hierarchies of features from inputs and is particularly suitable for problems that involve visual data. However, the performance of the CNN, largely depends on the design of the architecture and the process of training. Hence, many variations of enhancements are researched for the optimization of the CNN's performance [2, 3]. In this paper, an improved CNN architecture is proposed for the Fashion image classification problem. This is done with a significant emphasis placed on image augmentation and Batch Normalization techniques. The training dataset size can be increased artificially, most effectively, and easily by using image augmentation. These methods enable one to apply rotations, translations, zooms, flipping, and other transformations to the images. Image augmentation prevents overfitting and boosts the classifier's invariance by training the model on more variations of images [4, 5]. Another important addition is Batch Normalization, which tackles the so-called internal covariate shift problem in neural networks. By normalizing the layer inputs, Batch Normalization stabilizes the learning process, accelerates convergence, and, in many cases, leads to better overall performance. This technique has become a standard part of most state-of-the-art deep learning models due to its ability to improve training dynamics and the accuracy of models [6, 7]. In the proposed CNN architecture, image augmentation is used along with Batch Normalization to increase performance in Fashion-MNIST. This dataset's complexity, diversity, and widespread community support makes it an excellent choice for this study's experimental needs, providing a challenging and manageable dataset that helps rigorously evaluate the proposed models. Initially, there are several convolutional layers, and the filter size is increased in each layer. After each convolutional layer, a Batch Normalization layer is introduced to have stable learning. Then, there are the Max Pooling layers that reduce the feature map size, after which fully connected layers are used which end in a SoftMax output layer for classification. The use of dropout layers also helps in regularization so that the problem of overfitting does not arise. The model is trained with the Adam optimizer, which is very efficient and adaptive to the change in the learning rate. The loss function is a sparse categorical cross-entropy as this is a multiclass classification problem. The test set is evaluated for checking the accuracy and generalizing capability of the model. The experiments conducted demonstrate a significant improvement in classification accuracy, with the proposed enhancements being compared to a baseline CNN model. The augmented dataset and the inclusion of Batch Normalization contribute to a more robust and accurate classifier, highlighting the effectiveness of these techniques in fashion classification tasks.

II. RELATED WORKS

Previous works have explored various architectures and techniques to enhance CNN performance on the Fashion-MNIST dataset. Authors in [8] introduce a hybrid quantum-classical algorithm, QDCNN, which incorporates dilated convolution into quantum convolutional layers. The QDCNN extends the concept of dilated convolution to quantum neural networks to capture a larger context during the quantum convolution process while reducing computational cost. The architecture integrates quantum layers with classical layers, employing quantum circuits for convolutional operations. Authors in [9] put forward a hybrid model of CNNs and Long Short-Term Memory (LSTM) for classification of Fashion-MNIST images. CNN for spatial feature extraction and LSTM are used to manage sequential dependencies. The model has the convolutional, pooling, and LSTM layers, where the neural network is built and then optimized over time based on error backpropagation to set the weights of these layers. Authors in [10] propose a shallow convolutional neural network (SCNNB) with batch normalization for faster training convergence and improved classification accuracy. The SCNNB architecture consists of 2 convolutional layers followed by 2 max-pooling layers, a fully connected layer, and a SoftMax output layer. The use of small convolutional kernels and batch normalization helps in reducing time and space complexity. Authors in [11] use the classic LeNet-5 architecture (a convnet) to classify the images of the Fashion-MNIST dataset. The figures show convolutional layers, pooling layers, and fully connected layers of an architecture. This study intends to enhance the accuracy and efficiency of the model to classify fashion images by optimizing the classic LeNet-5 model. It does not provide any special optimization or modification concerning the LeNet-5 architecture. Authors in [12] utilize multiple CNN models for image classification (Fashion-MNIST). They train a few different CNN architectures and compare how they perform. The study also illustrates the necessity of model selection and ensemble techniques for enhancing classification capability and resilience. Authors in [13] employ CNNs to classify images of fashion styles. The methodology involves training several CNN architectures, utilizing convolutional layers, batch normalization, and residual skip connections to improve learning efficiency and accuracy. The research explores the impact of various hyperparameters, including activation functions, optimizers, learning rates, dropout rates, and batch sizes, on model performance.

III. METHODOLOGY

This section provides the methodology for designing the advanced model architecture of a CNN for fashion classification. The spectrum of the methodology points to the discussions held about the dataset, preprocessing of the images, augmentation methods, and modeling work with the CNN in this study, and justifies each of the components of the model.

A. Dataset

In this work, the dataset used is Fashion-MNIST, proposed by the Zalando Research lab. Fashion-MNIST is a more complex variant of their MNIST dataset and is made for evaluating image classification algorithms specific to fashion items. Fashion-MNIST contains 70,000 grayscale fashion images divided into ten different classes. It has classes such as T-shirts/tops, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots. As shown in figure 1, every image of the Fashion-MNIST dataset is 28×28 pixels in size. It is small and, hence, efficient in processing, yet big enough to allow for representative real-life difficulties regarding fashion items. It is already split into 60,000 images for training and 10,000 for testing; thus, model evaluation and comparison will cause no problem. Image preprocessing and augmentation would be two critical processes in preprocessing the dataset that shall be used to train the CNN model. This will normalize data and synthetically enlarge the size of the training set while enhancing the job ability of the models to generalize on new, unseen images.

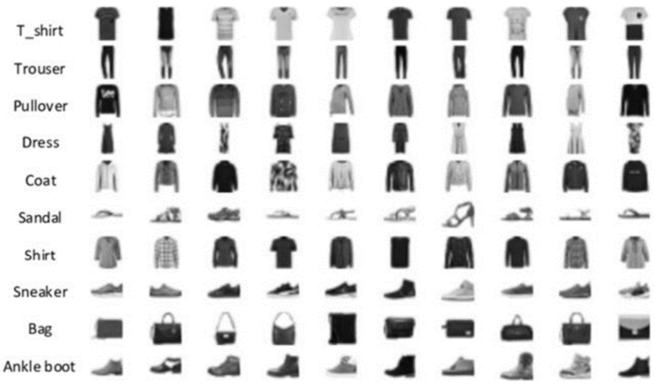

Fig. 1. Sample images from Fashion MNIST dataset.

B. Image Preprocessing

All the images are first normalized to be within a value range of [0, 1] for any amount of augmentation to be applied. Mainly, normalization is done by dividing each pixel by 255 so that the values can be constantly split between 0 and 1. This will, therefore, make data distribution be the same so that the neural network would quickly converge in its training phase [14].

C. Augmentation

Image augmentation is an effective tool used to artificially increase the size of the training dataset and the variability by randomly transforming the input images. It will not make the model biased and it will be robust against variation of input data [15]. The following augmentation techniques were applied:

- Rotation: Randomly rotates the image at a maximum of 10 degrees, within specified degrees of rotational orientation. This would allow the model to become invariant to rotational variations in images.
- Width and Height Shifts: A shift of up to 10% of the image dimensions will be randomly applied across both the horizontal and vertical dimensions. This will simulate the variation in positioning and slight misalignment of images.
- Zooming: Up to 10% of random zoom will be performed in and out of pictures. This way, the model learns to recognize objects at different scales.
- Horizontal and Vertical Flips: The image would be randomly flipped horizontally, as well as vertically. This can be useful with fashion items as they can be viewed in every orientation.
- Fill Mode: Now the transformation strategy could create empty areas in the image due to the transformations. Hence, the fill mode should be set to "nearest".

D. CNN Architecture

The designed architecture of the proposed CNN model is deployed for effective feature extraction and learning from the Fashion MNIST dataset after augmentation. The model is composed of several layers. Each of these layers serves specific purposes associated with the feature extraction process in any image classification task, as the one discussed here, as shown in Figure 2.

- Input Layer: The input layer accepts images of size 28×28 pixels with a single channel (grayscale). The input shape is specified as (28, 28, 1).
- Convolutional Layers: The proposed model uses 4 convolutional layers, followed by Batch Normalization layers, which serve to stabilize and accelerate the training process. Each convolutional layer uses Rectified Linear Unit (R.L.U) activation functions, a simple element-wise nonlinear function.
- Convolution Layer 1: The first layer with 32 filters of size 3×3 helps to learn low-level features such as edges and textures.

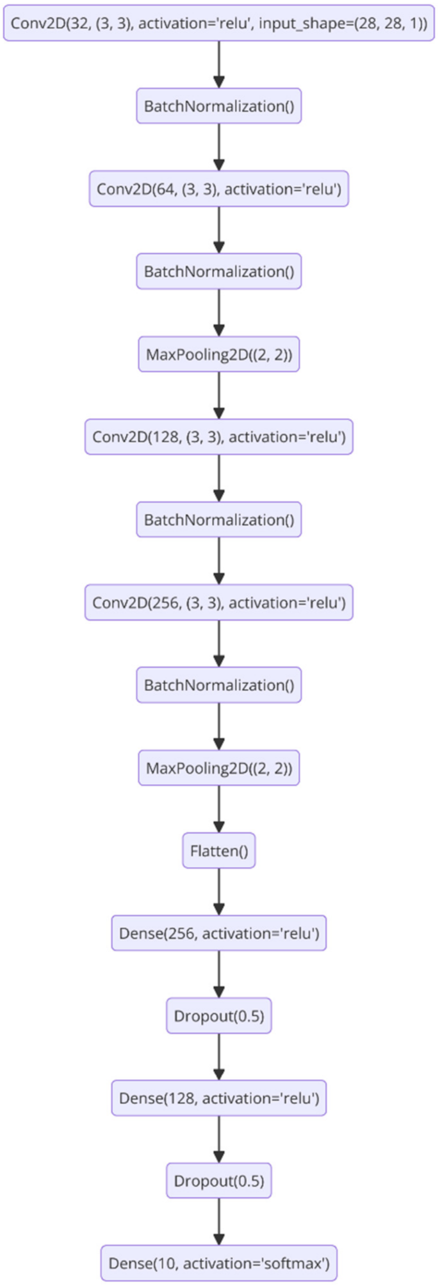
Conv2D(32, (3, 3), activation='relu', input_shape=(28, 28, 1))
BatchNormalization()
Conv2D(64, (3, 3), activation='relu')
BatchNormalization()
MaxPooling2D((2, 2))
Conv2D(128, (3, 3), activation='relu')
BatchNormalization()
Conv2D(256, (3, 3), activation='relu')
BatchNormalization()
MaxPooling2D((2, 2))
Flatten()
Dense(256, activation='relu')
Dropout(0.5)
Dense(128, activation='relu')
Dropout(0.5)
Dense(10, activation='softmax')

Fig. 2. Flowchart of Our Proposed Model.

- Convolution Layer 2: With 64 filters of size 3×3, this layer captures more varied and complex patterns based on the features learned in the first layer.
- Convolution Layer 3: This layer has 128 filters of size 3×3 to detect even finer features within the images.
- Convolution Layer 4: This layer has 256 filters of size 3×3, and it helps the model learn very complex features to accurately differentiate between different items of fashion.
- Each convolutional layer is followed by a Batch Normalization layer to normalize the output of the layer,

which tackles the issue of internal covariate shift. Stabilization improves the progress of learning, and higher learning rates can be utilized, leading to faster convergence.

- Max Pooling Layers: Max Pooling layers are used after the second and fourth convolutional layers to reduce the spatial dimensions of the feature maps. This helps in reducing the computational complexity of the model and prevents overfitting by progressively reducing the spatial size of the representation. The Max Pooling layers use a pool size of 2×2.
- Flatten Layer: The flattened layer itself automatically converges these 2D feature maps resulting from the last Max Pooling layer into a 1D feature vector. This step must always be done to push the data into the fully connected layers for classification.
- Fully Connected Layers: The model has two fully connected (dense) layers, each followed by a dropout layer to prevent overfitting.
- First Dense Layer: This layer has 256 units with R.L.U activation. It learns high-level representations from the flattened feature vector.
- Second Dense Layer: This layer has 128 units with R.L.U activation. It further refines the high-level features extracted by the previous layer.
- For each one of these dense layers, there is a dropout layer with a dropout rate of 50% right after. It helps to avoid overfitting by not overemphasizing every set of features.
- Output Layer: There are ten units in the output layer corresponding to the ten classes in the Fashion-MNIST dataset. The SoftMax activation function on this layer permits the building of a probability distribution over the classes, thus resulting in classification into several courses.

IV. EXPERIMENTS AND RESULTS

This section presents the results from training the developed CNN-enhanced architecture on the Fashion-MNIST dataset. Here, the results depict the model's accuracy for images with augmentation and Batch Normalization, as well as without them. The main idea of the experiments was to find how helpful image augmentation and Batch Normalization are for improving the CNN model's accuracy and generalization. This experiment was conducted on an Intel (R) Core (TM) i7-13620H 2.40 GHz RAM 16GB computer with an open-source library (Keras) that provides a Python interface for deep learning.

A. Training and Evaluation

This Training was carried out by feeding the augmented images to the CNN model for over 100 epochs, using Adam optimizer with a learning rate of 0.001 and a batch size of 32. The performance of the model was monitored against various metrics, such as training accuracy, validation accuracy, and test accuracy. The metrics to check were:

1. Training accuracy: how well the model fits training data.
2. Validation accuracy: how well the model performs with validation data (test set), giving an estimate of the generalization ability.

The main points of interest and test accuracy demonstrate how well the model has performed over unseen data. It was ensured that the values of training and validation loss were at a point, proving that the model is not overfitting on training data. The training accuracy rises gradually and hits around 95.76% toward the end, which is the corresponding drop in the training loss, indicating that the model learns effectively from the augmented data. This present learning process became stabilized and accelerated essentially through Batch Normalization. Validation accuracy behaved similarly and reached around 91.97% at the end of the training. Growth seems to be quite progressive, increasing gradually with the data. Thus, the decrease in validation loss was parallel to the training loss, proving proper learning with minor overfitting. Following this learning process, the model had a test accuracy evaluation performance of 91.97%. This will be highly supportive in the context of the complexity that characterizes the Fashion-MNIST dataset and will reinforce the implementation of improvements on the adopted CNN architecture.

B. Comparison with Baseline Model

To highlight the improvements achieved, the performance of the enhanced CNN model was compared with a baseline CNN model that did not incorporate image augmentation or Batch Normalization. The baseline model had the same architecture but was trained on the original dataset without any enhancements. The results provided in Table I exhibit that the enhanced CNN model outperformed the baseline model across all metrics. The test accuracy improvement from 88.5% to 91.97% is particularly noteworthy, demonstrating the practical benefits of the proposed enhancements.

TABLE I. COMPARISON BETWEEN ENHANCED AND BASELINE CNN

Model	Training Accuracy	Validation Accuracy	Test Accuracy
Baseline C,N.N	93.5%	88.2%	88.5%
Enhanced CNN (Proposed)	95.76%	91.97%	91.97%

C. Discussion

The experimental results validate the hypothesis that image augmentation and Batch Normalization significantly enhance the performance of CNN models for fashion classification tasks. Those techniques address some of the key challenges in training deep learning models, such as overfitting and internal covariate shift. Impact of Image Augmentation: By artificially increasing the diversity of the training data, image augmentation helps the model to learn more robust features. This is crucial for fashion classification, where items can appear in various orientations, scales, and positions. The augmentation techniques applied in this study (rotation, shifts, zoom, flips) effectively simulated these real-world variations, leading to improved generalization. Batch normalization normalizes the output of each layer, which can mitigate the problem of internal covariate shift and make the whole training process more stable and quicker. This could be achieved with a higher learning rate that would fasten the convergence rate and thus positively influence the overall performance. Here, it is reflected that the addition of Batch Normalization layers heavily affected gradual and steady increases in both training and validation accuracy. Limitation and Future Directions: This study is limited by its focus on one dataset the Fashion-MNIST, which may not generalize more diverse or real-world images. While the present study improves performance, there remain a few possible directions for future research. One may further look at integrating advanced techniques, such as attention mechanisms, to improve model performance. Applying these proposed enhancements to another fashion dataset is underway to establish the effectiveness of improvements deployed in different scenarios.

V. CONCLUSION

It has been proven that augmenting the images and using the Batch Normalization layer in the CNN architecture helps to get better results on the problem of fashion classification. The improved model has achieved better test accuracy, compared to the baseline model, at around 91.97%. Techniques such as rotation, shifting, zooming, flipping, and other augmentations not only help create data variation, but also significantly improve the model's generalization. Essentially, Batch Normalization contributed on making the learning process more stable, reducing internal covariate shift, and increasing convergence speed. Together, these techniques made overfitting smaller, which enables the model to learn more general features and, thus, outperform a traditional CNN. It was empirically witnessed that the conjecture that those techniques would be able to ameliorate a lot of CNN models for the fashion classification task could be made into a fact. Further studies could be undertaken on integrating more advanced techniques and applying these enhancements to other datasets to increase the validation of the methods. In summary, the above results suggest that robust data preprocessing and stabilization methods are essential for training efficient deep learning models in complicated image classification tasks.

REFERENCES

[1] S. Shirkhani, H. Mokayed, R. Saini, and H. Y. Chai, "Study of AI-Driven Fashion Recommender Systems," *SN Computer Science*, vol. 4, no. 5, Jul. 2023, Art. no. 514, https://doi.org/10.1007/s42979-023-01932-9.

[2] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, "Review of Image Classification Algorithms Based on Convolutional Neural Networks," *Remote Sensing*, vol. 13, no. 22, Jan. 2021, Art. no. 4712, https://doi.org/10.3390/rs13224712.

[3] O. M. Ahmed, L. M. Haji, A. M. Ahmed, and N. M. Salih, "Bitcoin Price Prediction using the Hybrid Convolutional Recurrent Model Architecture," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11735–11738, Oct. 2023, https://doi.org/10.48084/etasr.6223.

[4] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A Comprehensive Survey of Image Augmentation Techniques for Deep Learning," *Pattern Recognition*, vol. 137, May 2023, Art. no. 109347, https://doi.org/10.1016/j.patcog.2023.109347.

[5] A. Alzahrani, "Digital Image Forensics: An Improved DenseNet Architecture for Forged Image Detection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13671–13680, Apr. 2024, https://doi.org/10.48084/etasr.7029.

[6] J. Bjorck, C. Gomes, B. Selman, and K. Q. Weinberger, "Understanding Batch Normalization." arXiv, Nov. 30, 2018, https://doi.org/10.48550/arXiv.1806.02375.

[7] R. Ramesh and S. Sathiamoorthy, "A Deep Learning Grading Classification of Diabetic Retinopathy on Retinal Fundus Images with Bio-inspired Optimization," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 11248–11252, Aug. 2023, https://doi.org/10.48084/etasr.6033.

[8] Y. Chen, "QDCNN: Quantum Dilated Convolutional Neural Network." arXiv, Oct. 29, 2021, https://doi.org/10.48550/arXiv.2110.15667.

[9] Y. Ji, "A novel CNN+LSTM classification model based on fashion-MNIST," vol. 12258, Art. no. 122580S, Jul. 2022, https://doi.org/10.1117/12.2639667.

[10] F. Lei, X. Liu, Q. Dai, and B. W.-K. Ling, "Shallow convolutional neural network for image classification," *SN Applied Sciences*, vol. 2, no. 1, Dec. 2019, Art. no. 97, https://doi.org/10.1007/s42452-019-1903-4.

[11] M. Kayed, A. Anter, and H. Mohamed, "Classification of Garments from Fashion MNIST Dataset Using CNN LeNet-5 Architecture," in *2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)*, Oct. 2020, pp. 238–243, https://doi.org/10.1109/ITCE48509.2020.9047776.

[12] O. Nocentini, J. Kim, M. Z. Bashir, and F. Cavallo, "Image Classification Using Multiple Convolutional Neural Networks on the Fashion-MNIST Dataset," *Sensors*, vol. 22, no. 23, Jan. 2022, Art. no. 9544, https://doi.org/10.3390/s22239544.

[13] A. Vijayaraj *et al.*, "Deep Learning Image Classification for Fashion Design," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, 2022, Art. no. 7549397, https://doi.org/10.1155/2022/7549397.

[14] B. Ruff, T. Beck, and J. Bach, "Mean Shift Rejection: Training Deep Neural Networks Without Minibatch Statistics or Normalization." arXiv, Nov. 29, 2019, https://doi.org/10.48550/arXiv.1911.13173.

[15] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, Jul. 2019, Art. no. 60, https://doi.org/10.1186/s40537-019-0197-0.