A Hybird Framework based on Autoencoder and Deep Neural Networks for Fashion Image Classification

Aziz Alotaibi
College of Computers and Information Technology, Computer Science Department
Taif University, Taif 21974, Saudi Arabia

***Abstract*—Deep learning has played a huge role in computer vision fields due to its ability to extract underlying and complex features of input images. Deep learning is applied to complex vision tasks to perform image recognition and classification. Recently, Apparel classification, is an application of computer vision, has been intensively explored and investigated. This paper proposes an effective framework, called DeepAutoDNN, based on deep learning algorithms for apparel classification. DeepAutoDNN framework combines a deep autoencoder with deep neural networks to extract the complex patterns and high-level features of fashion images in supervised manner. These features are utilized via categorical classifier to predict the given image to the right label. To evaluate the performance and investigate the efficiency of the proposed framework, several experiments have been conducted on the Fashion-MNIST dataset, which consists of 70000 images: 60000 and 10000 images for training and test, respectively. The results have shown that the proposed framework can achieve accuracy of 93.4%. In the future, this framework performance can be improved by utilizing generative adversarial networks and its variant.**

***Keywords*—Fashion detection; fashion classification; convolutional autoencoder; deep learning; insert**

I. Introduction

Conventional machine learning algorithms still used in the image processing and computer vision to perform data mining and feature extraction such as support vector machine (SVM). However, conventional machine learning algorithms have some limitations when dealing 1) with unstructured large-scale data and 2) the utilization of available advance computer resources with graphic processing unit (GPU) and tensor processing unit (TPU) [1]. With the great advancement in deep learning techniques exploring and solving the most complicated computer vision tasks, many computer vision applications have gained a significant attention due to the available resources and large amount of data. Therefore, these factors allow for the acceleration of models' training and for alleviating of the vanish gradient through creating a deep learning model such as ResNet50. In addition, deep learning algorithms directly process raw data (e.g., an RGB image) and obviate the need for preprocessing phase and domain experts. The aim of feature engineering is to learn useful representations in order to make a better decision. Deep learning algorithms is divide into two types: discriminative learning algorithms and generative learning algorithms. In this study, we focus only on discriminative learning algorithms such convolution neural networks (CNN), recurrent neural networks (RNN), Neural network (NN), and a few to name. In the past few years, deep learning have been intensively searched for visual data processing including image classification, image detection, semantic segmentation, and video processing [2] such apparel classification. More recently, with the great development in online shopping and e-commerce, visual fashion analysis and recognition has attracted many researches to utilize both computer vision and deep learning techniques. Apparel detection and classification is becoming one of the most used computer vision applications in online industry for some advantages: first, online shop recognition and recommendation of desirable fashion products [3-5]. Second, enhancing user experience [6, 7]. Third, improving online advertising [8, 9]. Finally, improving the performance of human detection and recognition in different scenarios [10, 11]. However, apparel fashion applications still encounter some difficulties to perform well such as; 1) the variation of the taken image size, 2) the light, 3) the angle, 4) new style, 5) undefined subcategories, 6) . Furthermore, due to the complex pattern of fashion apparel, various fashion properties, and the performance of previous existing deep learning algorithms, we believe that applying a good representation learning technique to extract useful features would enhance the performance of the fashion apparel detection and classification. Therefore, we propose a novel framework utilizing deep spatial autoencoder. Therefore, the proposed framework has proven to extract robust representations and effective in classifying fashion images with accuracy of 93.4%.

The main contribution of this study can be summarized as follows:

1) Propose a novel fashion classification framework based on deep learning techniques.

2) Utilize the deep autoencoder to reduce the dimensionality of the input fashion image and deep neural networks as classifier.

3) The proposed framework achieves an accuracy of 93.4% on Fashion MNIST which outperforms some of the existing deep learning algorithms such as CNN.

4) The proposed framework are evaluated through several experimental analyses.

The main contribution of this work is that it proposes a novel framework for fashion classification. The remainder of this paper is organized as follows: Section 2 discusses the previous work related to the proposed framework. In Section 3, the proposed framework including both deep autoencoder and deep neural networks are demonstrated and explained. Section 4 illustrates and discusses the evaluation performance. Finally, Section 5 concludes this work and introduces future works.

II. Related Work

In this section, a literature review is conducted and only covered the most recent related studies that mainly employee the machine learning and deep learning algorithms on Fashion MNIST dataset and its applications. Soe at el. [1] proposed a Hierarchical Convolutional Neural (H–CNN) for fashion image classification which emphasize on the knowledge embedded classifier outputting hierarchical information. The H-CNN is utilizing both VGG16 and VGG19 as base model which composed of five convolutional layers, max-pooling layers and fully connected layers for both feature extraction and classification. In addition, Duana et al. [12] utilized VGG-11 model which consists of convolution layers, maximum pooling layers and followed by batch normalization layers to classify Fashion images. In [13], authors presented a deep convolutional neural network (ConvNet) to classify the fashion images. In addition, author in [14] proposed a CNN model with Support Vector Machine for Fashion images classification. The author in [15] introduced various Hyper-Parameter Optimization (HPO) methods and regularization techniques using four deep convolution layers for recognizing images of fashion objects. Bhatnagar et al. [7] presented a three different deep convolution network with batch normalization and residual skip connection to acceleration the learning process. Their model reports enhanced accuracy of around 2% over the other deep learning systems. Authors [11, 16] proposed a deep Capsule Network, DeepCaps, which uses the concepts of skip connections and 3D convolutions. The skip connection allows a good gradient within a Capsule cell during the backpropagation optimization. Authors utilized the novel dynamic routing algorithm to assist the learning process of deep Capsule network. In addition, Deliege et al. [17] introduced a deep learning networks, called HitNet, with capsules embedded for data augmentation. HitNet uses a hybrid Hit-or-Miss layer to synthesis the representative images of a specific class by utilizing a reconstruction network. In [18], authors presented a fashion images classification system using single feature descriptor, histogram of oriented gradient (HOG), and multiclass support vector machine (SVM) for classification and detection of fashion images utilizing Fashion MNIST dataset. Shen [19] utilized the Long Short Term Memory (LSTM) to construct a model that can perform a classification task on Fashion MNIST dataset. Li et al. [20] proposed a personalized representation learning that is used for transforming shared feature extractors into personalized feature extractors. Their proposed study is based on two variants of Collaborative neural networks: Unconditional Collaborative Neural Networks (U-CoNN) and Conditional Convolutional Neural Networks (C-CoNN) for fashion image classification. In addition, Xiao et al. [21] introduced a multi-class object detection Coprocessor by combining Histogram of Oriented Gradient (HOG) feature and Local Binary Pattern (LBP) feature with a weighted Softmax classifier for image detection task such as Fashion MNIST.

III. Proposed Methodology

The aim of this proposed framework is to enhance the performance of the fashion classification. The proposed DeepAutoDNN framework based on deep learning techniques for fashion classification is illustrated in Fig. 1. The framework consists of three parts: First, deep autoencoder, second, deep neural networks and finally the classifier.

Algorithm 1

1. Input: training image set $\{x_i\}_1^n$

2. Deep Autoencoder computation
- A. Initialization of variable
- B. Compute encoder and decoder using Equation (1)(2)
- C. Minimize the construction error using Equation (3)
- D. Compute the latent representation $\{y_i\}_1^n$and update the weights
- E. Repeat step B, C, and D until convergence
- F. return $\{y_i\}_1^n$

3. Deep neural network
- A. input: the output of $\{y_i\}_1^n$
- B. train deep neural networks
- C. minimize the training error via backpropagation

4. softmax classifier
- A. predicted label y using weight in step 2 and 3

A. Deep Autoencoder

Autoencoder is an unsupervised representation learning algorithm based on artificial neural networks that composed of three typical layers; input layer, hidden layer and output layer [22]. The key goal of the autoencoder is to identify and disentangle the underlying hidden representation for a set of data. The process training of autoencoder consists of two types: encoder and decoder as shown in Fig. 2. The goal of the encoder is to map the input data to hidden representation, and the decoder is to reconstruct the original data from the hidden representation. The set of data is represented as $\{x_n\}_{n=1}^N$, where $X_n \in \mathbb{R}^m$, h_n donates the hidden representation, and $\hat{x}$ represents the reconstructed output. The encoder and decoder are calculated in equation 1 and 2 respectively as follows:

$$h_n = f\,(W_1 x_1 + b_1) \quad (1)$$

$$\hat{x} = g\,(W_2\, h_n + b_2) \quad (2)$$

Where f and g is the encoder and decoder function respectively, and W_1, W_2 donates the weight matrix of the encoder and decoder respectively.

The parameter sets of the autoencoder are optimized to minimize the reconstruction error:

$$(W_1, W_2) = \sum_{j \in \Omega_i} L(x, \hat{x}) \quad (3)$$

Where L donates the reconstruction error [23].

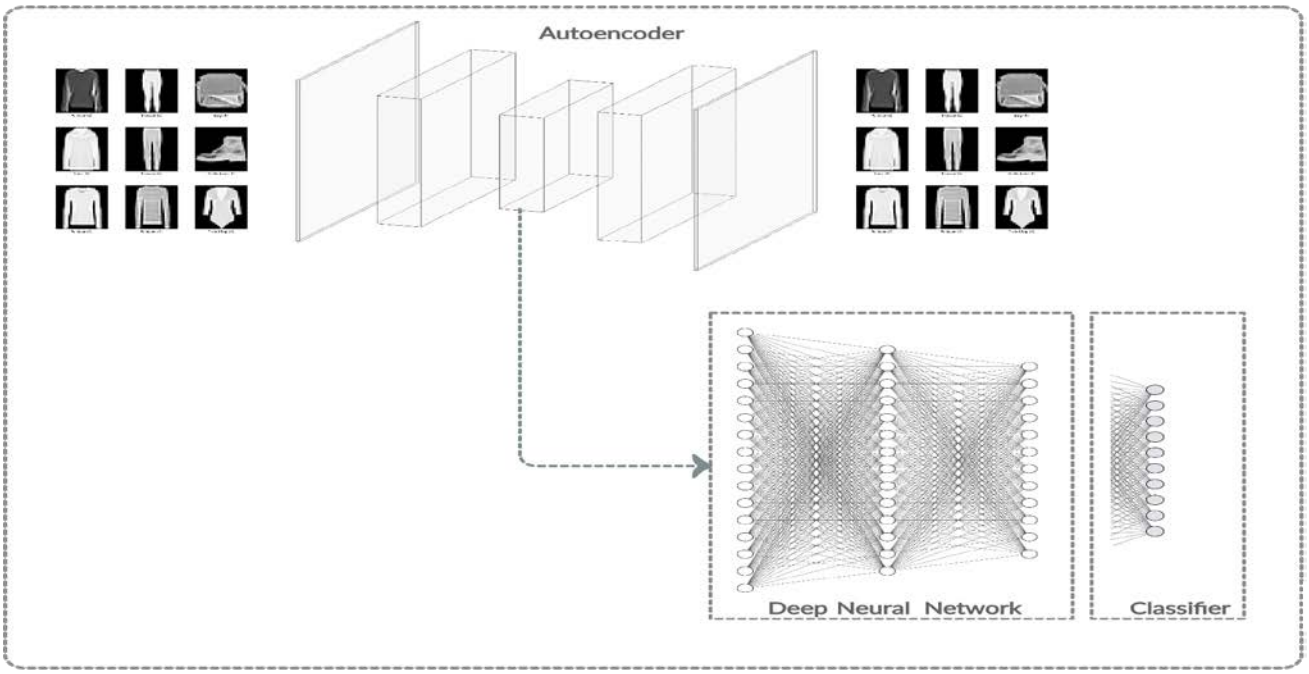


Fig. 1. The Overview Architecture of the Proposed Deep Autoencoder and Deep Neural Networks Framework.

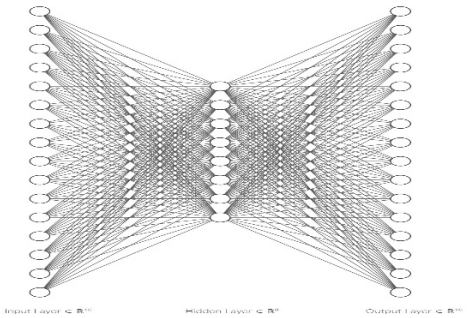


Fig. 2. Typical Autoencoder.

The proposed deep autoencoder consists of decoder and encoder as shown in Table I. The decoder is designed based on ConvNet2D layers followed by BatchNormalization layer and MaxPooling layers, whereas the decoder is based on ConvNet2D followed by UpSampling2D.

B. Deep Neural Networks

Neural networks are a supervised learning that consists of multiple layers to discover underlying feature between different variables as shown in Table III. In this study, deep neural networks take the output of the hidden layer of the spatial autoencoder 7 * 7 layer as an input which was flatten. The proposed deep neural networks consist of eight layers as shown in equation 4.

$$h_n = g\,(W_1 x_1 + b_1) \quad (4)$$

Furthermore, rectified linear unit activation function (ReLU) is applied to allow our framework to easily get sparse representations [24]. ReLU prunes the negative value to zero, and keeps the positive number x to the same value as shown in equation 5 and Fig. 3.

$$(x) = max(x, 0) \quad (5)$$

In addition, dropout layer is regularization technique that used to prevent/reduce overfitting issue especially when using huge networks [25, 26]. In our framework, dropout layer is introduced in both autoencoder and deep neural network to avoid the overfitting problem.

C. Classifier

The last layer of this proposed framework is the classifier layer with 10 neurons to classify the output into one of the fashion classes. The softmax activation layer is used to output the probability distribution for categorical classification and defined as follows [27]:

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_i}} \quad (x)$$

Where, x_i is the dimension of the input vector, and it will return a value between zero and one for each class and 1 for all classes.

TABLE I. DEEP AUTOENCODER ARCHITECTURE

Specific Parameter	Value/Type
Networks	Decoder and Encoder
Input shape	(28,28,1)
Epochs	15
Batch-size	32
Optimizer	Adam
Loss Function	mean_squared_error
Activation Function	Rectifier Linear Unit (ReLU) And Sigmoid
Pooling	MaxPooling and UpSampling

TABLE II. CLASS NAMES AND EXAMPLE IMAGES IN FASHION-MNIST DATASET

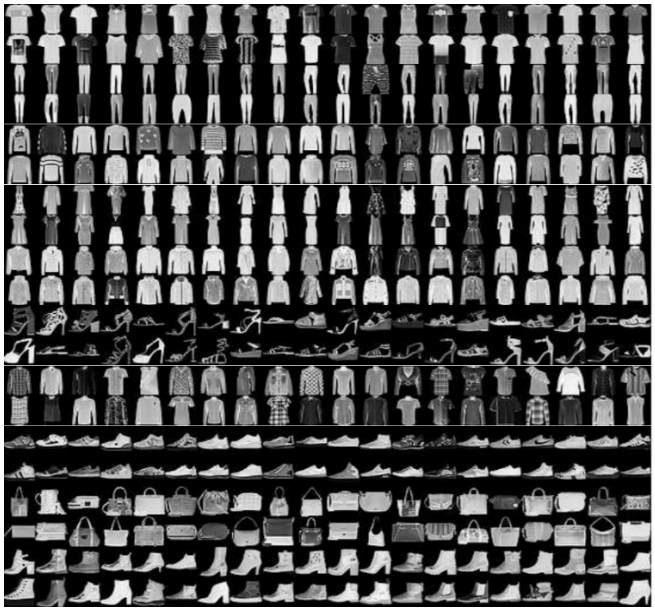
Label	Description	Example
0	T-shirt/top	
1	Trouser	
2	Pullover	
3	Dress	
4	Coat	
5	Sandal	
6	Shirt	
7	Sneaker	
8	Bag	
9	Ankle boot	

TABLE III. DEEP NUERAL NETWORK ARCHITECTURE

Specific Parameter	Value/Type
Layers	7
Input shape	(14, 14, 128)
Epochs	120
Batch-size	64
Optimizer	Adam
Learning Rate	0.001
Loss Function	categorical_crossentropy
Activation Function	ReLU
Dropout	0.5
Last Layer	Softmax

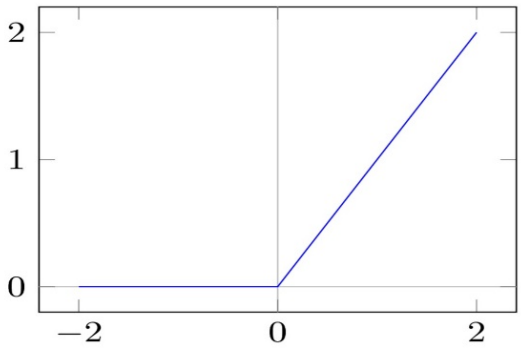


Fig. 3. Rectifier Linear Units (ReLU).

IV. DISCUSSION AND EXPERIMENTAL RESULTS

The goal of this section is to demonstrate the performance of the proposed framework based on deep autoencoder and deep neural networks. The proposed framework is tested on Fashion MNIST dataset. This section is divided into three subsections: (1) Fashion MNIST Dataset, (2) Performance evaluation, and (3) Discussion.

A. Fashion-MNIST Dataset

Fashion-MNIST [28] was introduced in 2017 and considered as replacement of MNIST dataset. This dataset composes of ten categories: t-shirt, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag and ankle boot. Fashion-MNIST consists of 70000 images which divided as follows: 60000 and 10000 for training and test respectively. each category has 6000 and 1000 images for training and test respectively. Images are grayscale with size of 28 by 28 as shown in Table II.

B. Performance Evalution

In this subsection, the classification performance of the proposed Autoencoder framework is evaluated using the Fashion MNIST dataset. We compared the performance of our proposed DeepAutoNN framework with some of the previous proposed approaches in term of overall accuracy. The compared approaches were: Hierarchical Convolutional Neural based on VGG16 and VGG19 [1], VGG11 [12], Convolutional Neural Network with Support vector Machine [14], Convolutional Neural Network followed by Batch Normalization and Skip Connection [7], HitNet [17], Histogram of Oriented Gradient with Multiclass Support Vector Machine [18], Long-Short Term Memory [19], Unconditional and conditional Collaborative Neural Networks [20], Histogram of Oriented Gradient with Local Binary Pattern [21].

As shown in Table IV, the proposed framework achieves the best result with accuracy of 93.4% compared to other proposed approaches.

TABLE IV. PERFORMANCE COMPARISON WITH THE PROPOSED SYSTEM

Model	Accuracy
H-CNN using VGG19 [1]	93.33%
VGG-11 [12]	91.5%
CNN-SVM [14]	90.72%
CNN2 + BatchNorm + Skip [7]	92.54%
HitNet [17]	92.30%
HOG + SVM [18]	86.53%
LSTM [19]	88.26%
CNN + U-CoNN [20]	90,61%
CNN + C-CoNN [20]	90,84%
ConvNet [13]	91.00%
HOG + LBP [21]	86.20%
DeepAutoNN-ours	93.36%

Confusion metrics are computed to estimate the classifier performance, and they are calculated as follows.

$$Accuracy\ (Acc) = (TP + TN)/(TP + TN + FP + FN)$$

$$Sensitivity\ (\text{Sen}) = \text{TP} / (\text{TP} + \text{FN})$$

$$Precision\ (\text{Pre}) = TP / (\text{TP} + \text{FP})$$

$$\text{Specificity (Spe)} = \text{TN} / (\text{TN} + \text{FP})$$

$$F - Score\ (F1) = (2 \times (Sen \times Pre) / (Sen + Per))$$

Where TP and TN represents the numbers of true-positive results, and true-negative and, respectively, and FN and FP denote the numbers of false-negative and false-positive results, respectively. The performance evaluation of the proposed framework using the evaluation metrics is shown in Table V. Furthermore, confusion matrices are used to further evaluate the multi-classification performance as shown in Fig. 4.

DATASET

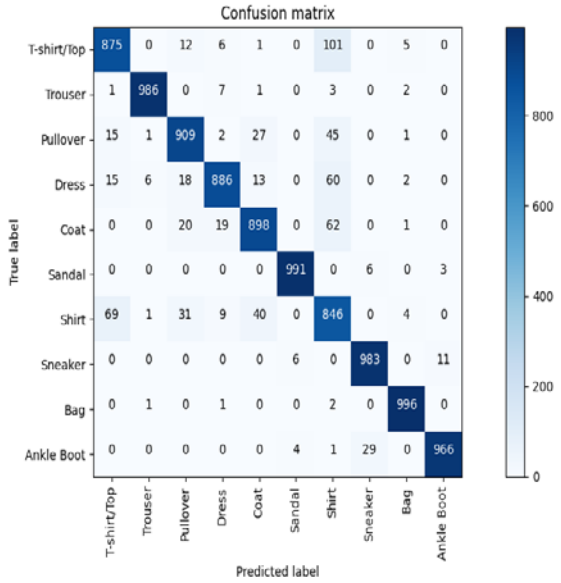


Fig. 4. Confusion Matrix of Classification result with Fashion MNIST.

C. Discussion

In this subsection, the efficiency of the proposed AutoDeep classification framework utilizing fashion images is discussed and analysis. The proposed deep autoencoder is used to extract the hidden and useful features to enhance the classification performance. We found out that using large number of feature maps assess the deep neural networks to converge easily, also, the softmax classifier performs well on the Fashion MNIST. Moreover, increasing the number of hidden layers of neural networks leads to better performance. The proposed framework converges fast as shown in Fig. 5 and Fig. 6 due to the reduce dimensionality using the autoencoder.

The computational time required by DeepAutoNN framework is further analysed which can be divided into three stages: deep autoencoder, deep neural networks and multi-classifier layer. The total time required for the DeepAutoNN framework to process and classify a single image is approximately 0.001 second/image. The proposed DeepAutoNN framework for fashion classification is implemented using Windows 10 with an Intel i7 CPU and 32 gb RAM. TensorFlow, Numpy, sklearn, matplotlib and the Keras library are utilized as tools to implement the framework.

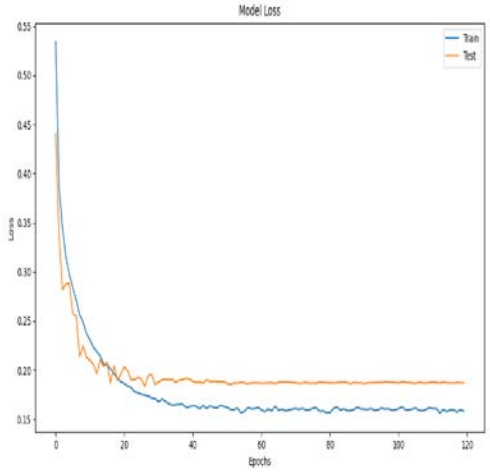


Fig. 5. Training Loss.

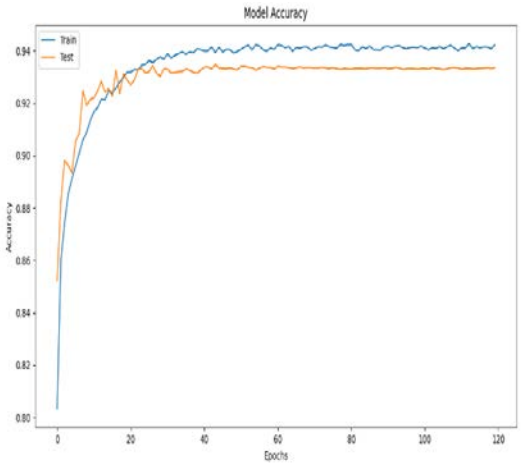


Fig. 6. Training Accuracy.

TABLE V. THE PERFORMANCE EVALUATION USING EVALUATION METRICS

Label	precision	Recall	F1-score
T-shirt/top	90	88	89
Trouser	99	99	99
Pullover	92	91	91
Dress	95	89	92
Coat	92	90	91
Sandal	99	99	99
Shirt	76	85	80
Sneaker	97	98	97
Bag	99	100	99
Ankle boot	99	97	98

V. CONCLUSION

With rapid advancement in deep learning and computer vision techniques, many studies have been conducted on complex vision tasks due to the powerful representation-learning algorithms utilizing advance deep learning techniques. Image detection and classification have been utilized in e-commerce industries such as apparel applications. This paper has proposed a novel framework that can detect and classify fashion images utilizing deep learning techniques. This framework combines a deep autoencoder with deep neural networks to extract the complex patterns and high-level features of fashion images. The framework AutoDeepDNN, demonstrates significant improvements over existing approaches on Fashion MNIST dataset with accuracy of 93.4% %. The future research plan is to explore and investigate generative adversarial networks and its variant to improve the classification results using more fashion datasets.

REFERENCES

[1] Seo, Y. and K.-s. Shin, Hierarchical convolutional neural networks for fashion image classification. Expert Systems with Applications, 2019. 116: p. 328-339.

[2] Pouyanfar, S., et al., A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR), 2018. 51(5): p. 1-36.

[3] Dong, Q., S. Gong, and X. Zhu. Multi-task curriculum transfer deep learning of clothing attributes. in 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). 2017. IEEE.

[4] Liu, Z., et al. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[5] Liang, X., et al. Human parsing with contextualized convolutional neural network. in Proceedings of the IEEE international conference on computer vision. 2015.

[6] Hadi Kiapour, M., et al. Where to buy it: Matching street clothing photos in online shops. in Proceedings of the IEEE international conference on computer vision. 2015.

[7] Bhatnagar, S., D. Ghosal, and M.H. Kolekar. Classification of fashion article images using convolutional neural networks. in 2017 Fourth International Conference on Image Information Processing (ICIIP). 2017. IEEE.

[8] Bossard, L., et al. Apparel classification with style. in Asian conference on computer vision. 2012. Springer.

[9] Yamaguchi, K., T.L. Berg, and L.E. Ortiz. Chic or social: Visual popularity analysis in online fashion networks. in Proceedings of the 22nd ACM international conference on Multimedia. 2014.

[10] Han, X., et al. Learning fashion compatibility with bidirectional lstms. in Proceedings of the 25th ACM international conference on Multimedia. 2017.

[11] Nair, P., R. Doshi, and S. Keselj, Pushing the limits of capsule networks. Technical note, 2018.

[12] Duan, C., et al. Image classification of fashion-MNIST data set based on VGG network. in Proceedings of 2019 2nd International Conference on Information Science and Electronic Technology (ISET 2019). International Informatization and Engineering Associations: Computer Science and Electronic Technology International Society. 2019.

[13] Elsaadouny, M., J. Barowski, and I. Rolfes. ConvNet Transfer Learning for GPR Images Classification. in 2020 German Microwave Conference (GeMiC). 2020. IEEE.

[14] Agarap, A.F., An architecture combining convolutional neural network (CNN) and support vector machine (SVM) for image classification. arXiv preprint arXiv:1712.03541, 2017.

[15] Greeshma, K. and K. Sreekumar, Hyperparameter Optimization and Regularization on Fashion-MNIST Classification. International Journal of Recent Technology and Engineering, 2019.

[16] Rajasegaran, J., et al. Deepcaps: Going deeper with capsule networks. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[17] Deliege, A., A. Cioppa, and M. Van Droogenbroeck, Hitnet: a neural network with capsules embedded in a hit-or-miss layer, extended with hybrid data augmentation and ghost capsules. arXiv preprint arXiv:1806.06519, 2018.

[18] Greeshma, K. and K. Sreekumar, Fashion-MNIST classification based on HOG feature descriptor using SVM. International Journal of Innovative Technology and Exploring Engineering, 2019. 8: p. 960-962.

[19] Shen, S., Image Classification of Fashion-MNIST Dataset Using Long Short-Term Memory Networks.

[20] Li, N., Y. Sheng, and H. Ni. CoNN: Collaborative Neural Network for Personalized Representation Learning with Application to Scalable Task Classification. in 2019 International Conference on Computer, Information and Telecommunication Systems (CITS). 2019. IEEE.

[21] Xiao, Z., et al., A Multi-Class Objects Detection Coprocessor With Dual Feature Space and Weighted Softmax. IEEE Transactions on Circuits and Systems II: Express Briefs, 2020. 67(9): p. 1629-1633.

[22] Liu, G., H. Bao, and B. Han, A stacked autoencoder-based deep neural network for achieving gearbox fault diagnosis. Mathematical Problems in Engineering, 2018. 2018.

[23] Wang, W., et al. Generalized autoencoder: A neural network framework for dimensionality reduction. in Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2014.

[24] Glorot, X., A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. in Proceedings of the fourteenth international conference on artificial intelligence and statistics. 2011.

[25] Hinton, G.E., et al., Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.

[26] Wu, H. and X. Gu. Max-pooling dropout for regularization of convolutional neural networks. in International Conference on Neural Information Processing. 2015. Springer.

[27] Shen, W. and R. Liu, Tackling Early Sparse Gradients in Softmax Activation Using Leaky Squared Euclidean Distance. arXiv preprint arXiv:1811.10779, 2018.

[28] Xiao, H., K. Rasul, and R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.