Review on Existing Technique on Fashion Article Image Classification

Research-based ☒ Application-based □

Name	Tan Xi En	Lee Hao Jie	YikaninYit
Programme	CS	CS	CS
ID	1904058	1801544	1806487
Contribution	1/3	1/3	1/3

1. INTRODUCTION

In the domain of image classification there are many state of art models that have been proposed to solve different kinds of computer vision problems. Some well-known models are Lenet5, AlexNet, VGG, Resnet and Squeeze Network. These well known models have been made in comparison to well known datasets such as ImageNet. This is a dataset that contains 1 million images with different sizes of images. In theory an image classification problem can be solved with a state of art model with sufficient trained time.

Nowadays people buy commodities through e-commerce platforms. Especially in this COVID-19 pandemic most of the people choose to buy their clothes online. There are requirements to automate the process for the annotation of new fashion products. This can reduce the workload for people and improve the speed for e-commerce platforms to make these products online. Fashion-MNIST is a small dataset compared to ImageNet because the image size are 28*28 with grayscale image only. We will use the Fashion-MNIST dataset to analyze different model behavior on a simpler dataset compared to ImageNet.

In this paper , we will focus on three different types of convolutional neural network and imply experiments on the Fashion-MNIST dataset. The three models we choose to analyze are the Lenet5 model, SEResnet model, Skip-connection with batch normalization model. Lenet5 is the simplest model compared to other models. SEResnet is the resnet model with Sequeeze connection to decide the importance of channel and the last model is theCNN model which uses Batch Normalization and Skip Connection with only 2 convolution layers.

2. RELATED WORK

Although identifying a visual entity is easy for a human, it is difficult for a computer algorithm to extract meaningful information with human level accuracy. Machine learning techniques such as SVM or KNN had subpar performance as they depend on the quality of the dataset to make meaningful inference. The CNN model was then proposed because the architecture of a CNN is similar to that of the connectivity pattern of neurons in the human brain. Individual neurons respond to stimuli in the restricted region of the visual field. Therefore, it can learn highly abstract features and can identify objects efficiently.

According to Ahmand et.al(2020), there are some problems faced to perform the classification on the Fashion-MINST dataset. One of the problems is the costume can be easily distorted by the lengthening pattern. Moreover, the same costume may have a different design to make them look differ but conceptually under the same category or different category of costume but possess some common features. Also, the costume items are difficult to be recovered due to their robustness. In addition, the presentation of the costume in an image will affect the effectiveness of the classification, for example, the angle of photo taken, light, noise background, and the costume is wearing by a model or just a photo of the costume. Hence, an algorithm is essential to have high performance multi-classes fashion classification. Ahmad Anter et al. (2020) has proposed a CNN based LeNet-5 model to solve the challenging issues as the proposed model is not used for the MINST dataset.

The study of Ahmad Anter et al. (2020) has shown how great the CNN based Lenet-5 architecture is in multi-classes image classification on Fashion-MINST datasets. After the model is trained, (Ahmad, Hadeer & Mohammad 2020) get a high accuracy of 98.9% across all 10 trials for the test set. The loss during the validation phase was much higher than the loss during the training phase. Moreover, the model has achieved a high performance on precision, recall and F-measure for each metric. Not only that, LeNet-5 scores the highest accuracy than other CNN models.

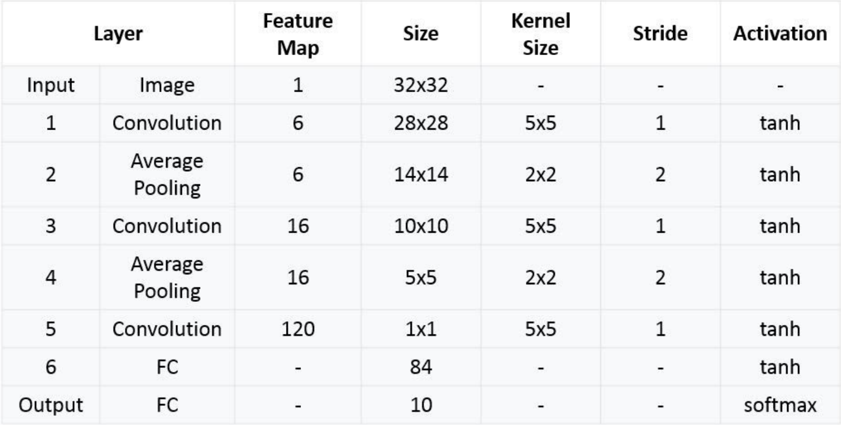
The Resnet model has shown its power of the residual and its performance on the ImageNet dataset. The result is significant; it beat other models on the ILSVRC 2015 competitions. Attention and gating mechanism has given Jie Hu et.al (2017) some inspiration because of its utility across many tasks, for instance, sequence learning, image captioning and lip reading. The result Jie Hu et.al(2017) proposed a SE block which is a light-weight gating mechanism to control the information passing after the residual block. Output of the gating mechanism is a set of channel weight modulation. These channel weights will combine with the feature map from residual block to generate a new feature map which has been recalibrated with the attention and gating mechanism.

According to the study of Jie Hu et.al(2017).SE block has made the model performance reduce around 1% of top 5 error rate in ImageNet dataset in residual network like Resnet and not residual network like VGG. The SE block was added after the residual block before skip connection with a reduction ratio of 16. Reduction ratio is a hyperparameter used to control the capacity and the computation cost of the two fully connected layers (Jie Hu 2017).SE block work by strengthening the class-agnostic feature in the earlier layer among all classes and show its specialise in later layer by respond differently to the inputs of high class-specific manner

The CNN model proposed by Bhatnagar, S., Ghosal, D., & Kolekar, M. H. was a 2 layered Conv Neural Network with Batch Normalization and Residual Skip Connection. The results of the model proposed above was that it achieved an accuracy of 92.54%. At the time period, CNN was a relatively new approach to image processing, therefore it performed much better than machine learning techniques. From the scoring metrics, the model proposed has 0.92 score for precision, 0.92 score for Recall and 0.92 score for F1 Score. The F1 Score is very high which represents low misclassification error. The top most misclassified fashion article items are T-shirt, Shirt, Pullover and Coat class. The author hypothesizes that the reason is because the dataset only contains grayscale images, which causes the images to look the same. Distinguishing features like color, texture are only available in RGB images, which is unfortunately not available in the fashion MNIST dataset. The author concludes that using a different dataset may reduce the misclassification error.

3. **SYSTEM DESIGN** (rename based on your method)

Lenet5



Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	32x32	-	-	-
1	Convolution	6	28x28	5x5	1	tanh
2	Average Pooling	6	14x14	2x2	2	tanh
3	Convolution	16	10x10	5x5	1	tanh
4	Average Pooling	16	5x5	2x2	2	tanh
5	Convolution	120	1x1	5x5	1	tanh
6	FC	-	84	-	-	tanh
Output	FC	-	10	-	-	softmax

Table 3.1 Lenet 5

The first model is built according to the architecture of the LeNet-5. The model consists of three convolution layers, two average pooling layers, and one fully connected layer. Each of the layers is followed by a tanh activation function and the stride of each convolution layer is set to one. The first layer, it reduces the size of the features map with size 5x5 and stride one. Second, a downsampling layer which is an average pooling using 2 strides to half the dimension of the receptive field. The third layer and fouth layer are same as frist two layer. In the fifth layer, a fully connected convolution layer is implied to have 120 features map. The last two layers will be the fully connected layer to have the final 10 features map for image classification. The cross-entropy activation will be implied to the last fully connected layer instead of softmax activation.

SEResnet

Layer Name	Output Size	SEResnet
stem	16×16×64	5x5 , 64 ,stride2
block1	16×16×64	$\begin{pmatrix} 3 \times 3,64 \\ 3 \times 3,64 \\ FC[4,64] \end{pmatrix} \times 2$
block2	8×8× 128	$\begin{pmatrix} 3 \times 3,128 \\ 3 \times 3,128 \\ FC[8,128] \end{pmatrix} \times 2$
block3	4×4×256	$\begin{pmatrix} 3 \times 3,256 \\ 3 \times 3,256 \\ FC[16,256] \end{pmatrix} \times 2$
global pooling	1×1×256	4×4 average pool
fully connected	10	256×10 fully connections

Table 3.2 SEResnet

Second model is a combination architecture from Resnet 18 and Squeeze net block. The network took a 32x32 size of grayscale image as input. To avoid the receptive filed become too small I made some modification on the resnet18 architecture. All convolution layer was followed by a batch normalisation layer and a ReLu activation. All blocks in the network first convolution layer use a stride of 2 to reduce the receptive field by half and double up the channel except the first block. After every two 3x3 convolution layers and SE modules a skip connection will be implied.

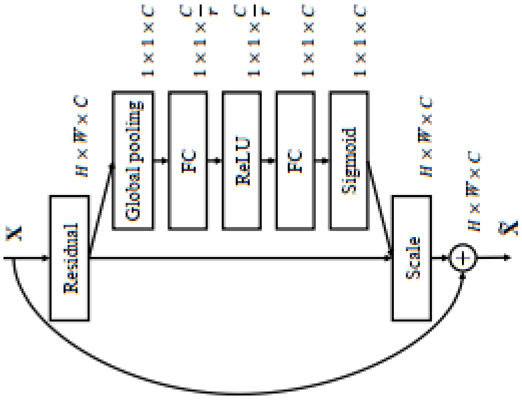
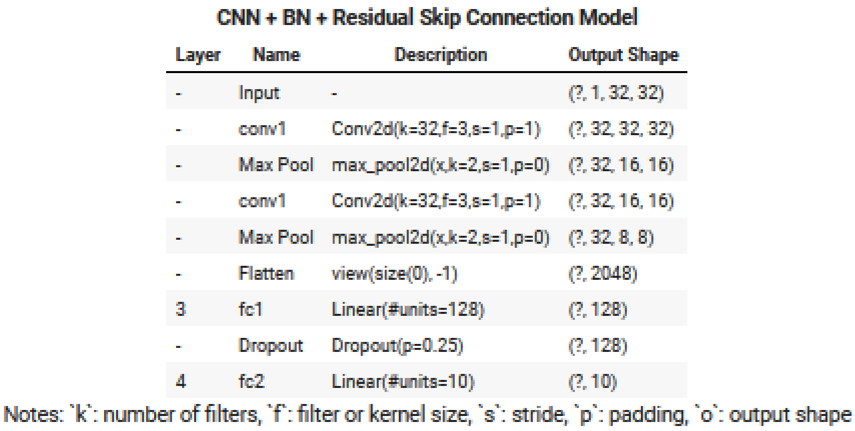


Figure 3.1 Squeeze Module

The Squeeze module is put after the residual module which is after two 3x3 kernels. C is the channels and r is the reduction rate which is 16. The output of the residual module will further to a global pooling layer to make a channel wise descriptor. Two Fully connected layers will act as the gating mechanism to decide what scale of each channel should retain. Sigmoid function is to make the output between range of 0 to 1 to scale the residual. The residual output will times with the scale factor feature from SE modules.

BN+skip connection



CNN + BN + Residual Skip Connection Model

Layer	Name	Description	Output Shape
-	Input	-	(?, 1, 32, 32)
-	conv1	Conv2d(k=32,f=3,s=1,p=1)	(?, 32, 32, 32)
-	Max Pool	max_pool2d(x,k=2,s=1,p=0)	(?, 32, 16, 16)
-	conv1	Conv2d(k=32,f=3,s=1,p=1)	(?, 32, 16, 16)
-	Max Pool	max_pool2d(x,k=2,s=1,p=0)	(?, 32, 8, 8)
-	Flatten	view(size(0), -1)	(?, 2048)
3	fc1	Linear(#units=128)	(?, 128)
-	Dropout	Dropout(p=0.25)	(?, 128)
4	fc2	Linear(#units=10)	(?, 10)

Notes: `k`: number of filters, `f`: filter or kernel size, `s`: stride, `p`: padding, `o`: output shape

Table 3.3 BN+Skip Connection

For the model proposed, it uses 2 convolutional and max pooling layers one after the other. Each convolutional layer has 32 filters of size 3x3 and max pooling which has filter size of 2 and stride size of 1.To improve the training speed of the model, Batch Normalization is done before every convolutional layer. During the training process, each convolution layer inputs are normalized by using the mean and variance of the values of the previous batch. This enables the network to train faster with higher learning rates. Residual skip connection is applied and used in the model to avoid vanishing gradient effect. In this model, we add the Batch normalized input and current value of convoluted output to get the final output.

After Batch Normalization and Skip Connection, the outputs of the convolutional layer are then flattened and passed into Dense or Fully Connected Layers. Both of the fully connected layers use ReLU as its activation function. The model also uses 25% dropout as a measure of regularization. This is to prevent overfitting issues in dense networks.

4. EXPERIMENT & EVALUATION

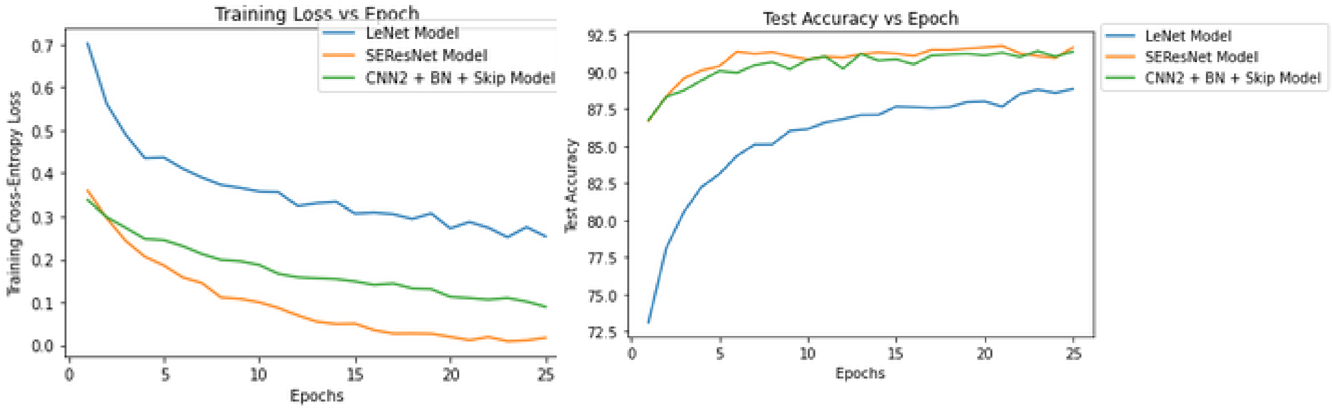
Model	Accuracy	False positive on shirt
Lenet5	88.85	303
SEResnet	92.08	230
BN+Skip connection	92.14	236

Table 4.1 Summary of Confusion matrix and classification report

Classification report and Confusion matrix

Based on the classification report, each model has achieved high accuracy in the classification of 10 classes, such as T-shirt, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, and ankle boot. The model two and model three have scored almost the same accuracy, 92.08% and 92.14% respectively. However, the model one only achieved accuracy of 88.85% which is lower than the others. This situation happened in an expectation as the model one is built by a simple architecture, LeNet-5. SEResnet was expecting higher accuracy than the BN+Skip connection model because it has more layers.

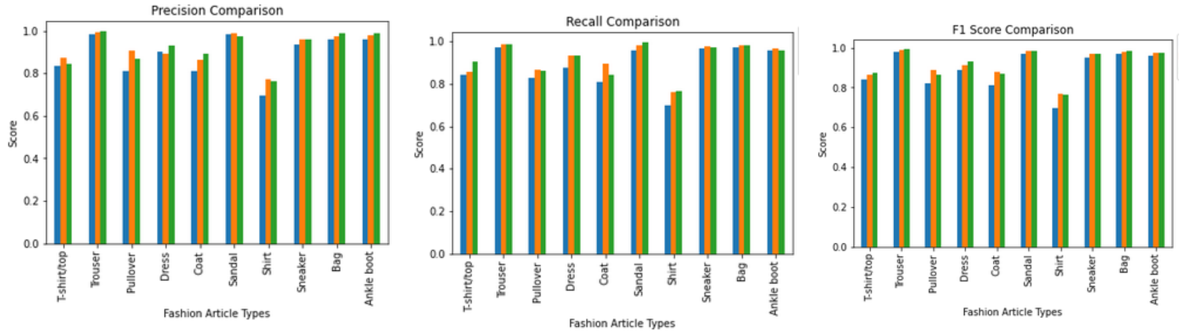
Although each model has achieved a high accuracy, each model has a high misclassification rate between T-shirt, pullover, dress, coat, and shirt. Each confusion matrix of each model has shown that the shirt has the highest misclassification rate between other classes. But the error seems to be decreased with the accuracy stated in the classification report. The highest accuracy model gets a lower misclassification rate. For example, classification on shirts using model one with accuracy 88.85% have 303 false positive on shirt, model two with accuracy 92.08% have 230 false positive on shirt, and model three with accuracy 92.14% have 236 false positive on shirt. The false positive on shirt occurred may be due to the similarity of the features of the classes.



Graph 4.1 Train loss vs Epoch and Graph 4.2 Test Accuracy vs Epoch

<u>Training loss across epoch and Test Accuracy vs Epoch</u>

Based on the graph of training loss across 25 epoch, we can observe that among the three models, LeNet-5 model has the lowest performance. It has a high training loss of 0.4 even after 25 epoch. SE ResNet Model has the best performance as the value of the training loss has the fastest descent to convergence. For the second graph, Test Accuracy across 25 epoch, SE ResNet model has the same accuracy as CNN + BN + Skip Model. Despite the difference in training loss, the test accuracy is the same compared between SE Resnet and CNN + BN + Skip model. LeNet-5 model still has the lowest accuracy.



Graph 4.3 Precision, Recall and F1 Score

<u>Percision recall , F1 , and accuarcy</u>

Accuracy of the model is not the only performance measure of the model. In multiclass-classes classification precision , recall and F1-score also can be used to compare performance of the model. Precision is the fraction of true sample within predicted sample and recall is fraction of how many true samples are predicted. For the purpose of applying the model in the Fashion Mnist dataset, precision is more important. For F1-score to be higher both precision and recall must be high. Model 1 has the lowest score among performance measures. The model 2 and model 3 do not have significant differences in each performance measure.

Visualization of Feature Maps

Visualizing CNN is often a great way to understand how a deep learning network learn new parameters and formulate features. For this assignment, we shall be looking and understanding the visualization of filters, feature maps and heatmap of Class Activation. The model we shall be using for visualization will be the third model, CNN + BN + Skip model.

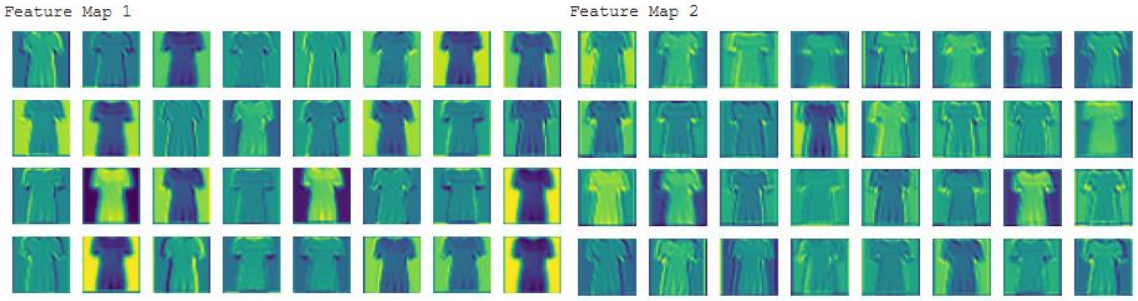


Figure 4.1 Feature Maps

From the 2 images above, we can observe that the T-shirt images in the second layer are much clearer, whereas the T-shirt images in the first layer are much blurrier. This is because at the 2nd convolutional layer, the model has begun selecting important features to identify T-shirt. We can observe that much clearer lines are drawn over the shoe giving it a shape, whereas the first layer has very blurry vision.

Visualization of Gradient Class Activation Map

While visualizing the feature map may provide us a lot of understanding, it is not sufficient as we do not know which features are focused on. Therefore, Grad-CAM approach is proposed to visualize which part of the images are focused on. By converting the gradient of the last convolutional layer into a 2d numpy array, we can plot it as a heatmap. The heatmap will show us which part of the image are focused on, like the feature maps.

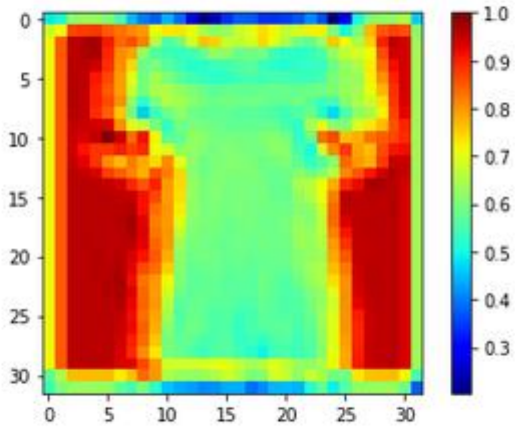


Figure 4.1 Heat Maps

From the above image, the color ranges from red => green => blue. The colors that approach red represent noise, whereas blue represents the features that are focused by the model. We can observe that the CNN model focuses on the fabric cloth of the image to classify it as a “T-shirt/top”. This provides us with good analysis, and a starting point to improve the classification power of our model.

5. CONCLUSION

The Lenet 5 model was considered outdated in the current time but it got an acceptable result with around 3% of difference on the accuracy. SEResnet expected to get high accuracy with its 12 layers of convolution with additional channel gating mechanism but it seems like overfit to this dataset even with trimming some convolution layers. Regarding the efficiency of training time and accuracy, model 3 is the best model for this dataset. In conclusion, a small size image dataset should consider simpler network architecture with some good modules such as skip connection to train efficiently. Our future work should focus on analyzing the effect of bigger size image fashion dataset to the efficiency of the model.

6. CONTRIBUTIONS

Contributions are listed in Fashion_MNIST_Model ipynb file.

Sources and Github Repo Used:
https://github.com/moskomule/senet.pytorch/blob/master/senet/se_module.py
https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/
https://medium.com/@stepanulyanin/implementing-grad-cam-in-pytorch-ea0937c31e82

Google Drive Link:
https://drive.google.com/drive/folders/1SCFhdqIMTEbKVldtK2dzOHchMVCGLxUa?usp=sharing

Github Repo (For This Assignment):
https://github.com/Neix20/Deep_Learning_May_2021

REFERENCES

Ahmed Anter, Mohammed Kayed and Hadeer Mohamed. (2020) “Classification of Garments from Fashion MNIST Dataset Using CNN LeNet-5 Architecture” in 2020 International Conference on Innovative Trends in Communications and Computer Engineering (ITCE’2020).

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

J. Hu, L. Shen and G. Sun, (2018) "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.