**COMP9444 Project Summary**

**Fashion Item Classification using Deep Learning**

**I. Introduction**

With the increasing demand for online shopping, there’s a need for accurate recommendation systems in e-commerce. The Polyvore dataset provides us a rich source of fashion item images and labels, making it ideal for developing these systems.

Currently, fashion item classification task face challenges like variety in styles, deformation and occlusion, which make it difficult to identify. The project aims to train and test different classification models to accurately detect clothing categories to find the most suitable model for this dataset.

**II. literature Review**

● SqueezeNet is a lightweight neural network designed to achieve high accuracy with minimal parameters, ideal for applications with limited computational resources. It achieves efficiency primarily through the Fire Module, which reduces the number of parameters while maintaining expressive power.

● MobileNetV2 introduces inverted residuals and linear bottlenecks, enhancing feature representation while significantly reducing computational cost, making it ideal for mobile devices. It demonstrates superior efficiency and accuracy across tasks..

● ViT show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks.

● Focal loss show the different loss to focus on the categories which has small samples.

● DenseNet is a deep learning architecture that emphasizes dense connections between layers, allowing each layer to receive input from all preceding layers, thereby facilitating feature reuse and improving gradient flow.

● ResNet model to effectively classify pneumonia cell images, demonstrating improved performance in medical image analysis by leveraging deep learning techniques

**III. Methods**

a. SqueezeNet is a lightweight convolutional neural network designed for image classification with significantly fewer parameters. It features fire modules, combining 1x1 and 3x3 convolutions, and a delayed downsampling strategy to retain high-resolution feature maps. Its compact size makes it easy to compute.

b. MobileNetV2 is a new neural network architecture designed specifically for mobile and resource-constrained environments, such as mobile phones and embedded devices. Compared to traditional neural networks, MobileNetV2 significantly reduces computational cost and memory usage while maintaining high accuracy. The key innovations in MobileNetV2 include the use of depth-wise separable convolutions, linear bottlenecks, and inverted residual connections.

c. ResNet50 was chosen as the backbone model because it proved to be effective in handling complex image classification problems. ResNet50's 50-layer architecture and residual connections enable the model to learn deep hierarchical features without being affected by the disappearance of gradients, making it ideal for capturing complex patterns in datasets. Its balance between depth and computational efficiency ensures that it can achieve high accuracy without requiring excessive computational resources.

d. DenseNet is a neural network architecture that can improve the efficiency of feature transfer, especially for environments with limited memory resources, such as mobile devices and embedded systems. Compared to traditional neural networks, DenseNet reduces computational costs and memory usage while maintaining high accuracy. DenseNet's key innovations include the introduction of dense connectivity, transition layers and bottleneck layers. The characteristic is that each layer is connected to all the layers before it, forming a dense connective structure, which improves the reuse of features and helps to communicate gradients and information more effectively. This dense connection can not only reduce the number of parameters, but also improve the training efficiency and convergence speed of the model.

e. In the Vision Transformer (ViT) model, an image is divided into 16x16 pixel patches, flattened into vectors, and passed through a linear projection layer for high-dimensional embeddings. Positional encodings and a classification token ([CLS]) are added to retain spatial information and serve as a global image representation for classification. The sequence, including the [CLS] token and patch embeddings, is fed into a Transformer encoder with multi-head self-attention and feed-forward layers, using residual connections and layer normalization for stability. After multiple encoder layers, the [CLS] token aggregates patch information and is passed to an MLP head for classification. To address the long-tail problem, focal loss is used to emphasize specific categories.

**IV. Experimental Setup**

**Model Optimizer Epoch ACC F1 lr weight decay scheduler**

**ViT 3e-04 1e-02 CosineAnnealing 50 0.86 0,84 MobileNetV2 1E-04 0 none 20 0.93 0.88 Densenet 1e-3 1e-4 none 20 0.85 0.84 SqueezeNet 1e-3 0 none 20 0.87 0.77 ResNet50 1e-3 1e-2 none 20 0.91 0.81**

a. Original Dataset

https://github.com/AemikaChow/AiDLab-fAshIon-Data/blob/main/Datasets/cleaned-maryland.md Processed Datasets (use the same train set and test set in each method)

https://drive.google.com/file/d/1yqNYJun2C3slsg2FXL7QW9ScMDgh8v08/view?usp=sharing b. Same Hyperparameters

Image dimensions: 224\*224 pixels

Number of classes: 20

All models are not pre-trained

c. SqueezeNet

The training process for SqueezeNet was conducted from scratch without pretrained weights,splitting into 80% as the training set and remaining as the test set.CrossEntropyLoss was used as the loss function, and the Adam optimizer managed the weight updates in the training test.

d. MobileNetV2

The model uses CrossEntropyLoss to compute the validation loss. During each epoch, it will be evaluated on the validation set after the model completes training on the training set. In the test section, ‘sklearn.metrics.f1\_score’ be used to get F1 score.

e. Resnet50

The experimental setup involved training a ResNet50 model for multi-class image classification, using a dataset split into 80% training and 20% testing subsets. Weighted cross-entropy loss and the AdamW optimizer were employed. Evaluation metrics included accuracy, macro-averaged F1, and weighted F1 scores to address class imbalance.

f. Densenet

The experimental setup includes various configurations to optimize model performance. RandomHorizontalFlip is applied to randomly flip images horizontally, while Normalization is performed using a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225] to match the input specifications of ImageNet pretrained models. For the Data Loader Configuration, a batch size of 32 is used for both training and testing. The training data loader is set to shuffle=True to ensure a different data order for each epoch, enhancing model generalization, while the test set order remains unshuffled. The Training Setup includes CrossEntropyLoss as the loss function, suitable for multi-class classification, and SGD (Stochastic Gradient Descent) as the optimizer with a learning rate of 0.001, momentum of 0.9, and a weight decay of 0.0001. The model is trained for 20 epochs to allow comprehensive feature learning.

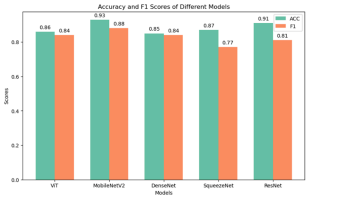
g. ViT

The training of the vit model is divided into two environments, one is based on the local GPU RTX 3090 CPU 7800X3D 32GB RAM, WINDOWS 10 Professional Edition environment, and the other is based on the colab A100 40GB environment. The two environments have the same parameters except that the batchsize of the training session is 256 and 512 respectively. For example, the batchsize of 32 is used for both valid and test. In the base model, only train and test dataset were used. when I compared the base model and the new model with focal loss, I used valid data extra, which was divided from 20% training data. So this part of the training valid and test set ratio is 16:4:5.

**V. Results**

In this project, we evaluated multiple deep learning models for fashion merchandise classification, each demonstrating unique strengths and limitations. MobileNetV2 has the highest accuracy (93.26%) and a Macro F1 Score(0.88), balancing efficiency and performance. ResNet50 achieved a high accuracy (91.04%) but performed poorly in minority forecasting, which is reflected in its low macro F1 score (0.81). While SqueezeNet and DenseNet offer lightweight architectures, their performance is mediocre due to depth and parameter redundancy limitations. vitb (Vision Transformer) introduces new transformer-based features, but requires a lot of computing resources. Overall, class imbalances remain a

key challenge, affecting minority class projections. Future work could explore integrated approaches and advanced data enhancements to improve robustness across all categories. The results of each method are showed in the corresponding notebook document.



**VI. Conclusions**

In this project, we evaluated multiple deep learning models for fashion merchandise classification, each demonstrating unique strengths and limitations. MobileNetV2 has the highest accuracy (93.26%) and a Macro F1 Score(0.88), balancing efficiency and performance. ResNet50 achieved a high accuracy (91.04%) but performed poorly in minority forecasting, which is reflected in its low macro F1 score (0.81). While SqueezeNet and DenseNet offer lightweight architectures, their performance is mediocre due to depth and parameter redundancy limitations. vitb (Vision Transformer) introduces new transformer-based features, but requires a lot of computing resources. Overall, class imbalances remain a key challenge, affecting minority class projections. Future work could explore integrated approaches and advanced data enhancements to improve robustness across all categories.

Future work could focus on adapting the model for real-time recognition, enabling it to classify fashion items in dynamic environments, such as live video streams or mobile applications. This would require optimizing the model's efficiency to handle continuous inputs and deliver rapid, accurate predictions, making it suitable for real-world applications in retail or interactive fashion experiences. Another promising direction is to integrate multimodal data, combining visual information with other data types like text descriptions, user preferences, or contextual information. This approach could enhance classification accuracy and provide a more comprehensive understanding of fashion items, aligning with personalized recommendations or enriched shopping experiences.

**VII. references**

[1] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).

[2] Iandola, F.N., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

[3] Çınar, A., Yıldırım, M. and Eroğlu, Y., 2021. Classification of pneumonia cell images using improved ResNet50 model. Traitement du Signal, 38(1), pp.165-173.

[4] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[5] Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[6] Ross T Y, Dollár G. Focal loss for dense object detection[C]//proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2980-2988.