Deep Learning CNN and Real-Time Recognition for Fashion Clothing Classification

*CS5330 Computer Vision Final Project

Srikanth Bonkuri
Khoury College of Computer Science
Northeastern University
Boston, MA
bonkuri.s@northeastern.edu

Eileen Chang
Khoury College of Computer Science
Northeastern University
Boston, MA
chang.ei@northeastern.edu

Karthik Komati
Khoury College of Computer Science
Northeastern University
Boston, MA
komati.k@northeastern.edu

Prateep Rao Malyala
Khoury College of Computer Science
Northeastern University
Boston, MA
malyala.p@northeastern.edu

***Abstract*—Image classification, which seeks to identify categories of objects or scenes in images, is a supervised machine learning problem that has a variety of applications within the field of computer vision. Fashion classification, in particular, aims to identify categories of clothing items, and this appeals to emerging applications, such as e-commerce, computer-aided fashion design, social media, and video surveillance. In this paper, the Fashion MNIST and DeepFashion data sets are used as the basis for training and developing deep convolutional neural networks for fashion clothing classification. The trained deep networks are then used to implement a real-time clothing recognition system.**

***Index Terms*—fashion mnist, real-time recognition, classification, deep learning, convolutional neural network, deep fashion**

I. Introduction

Fashion-MNIST is a standard dataset of clothing article images and consists of a training set of 60,000 28x28 grayscale images and a test set of 10,000 images. Ten types of clothing, such as t-shirts, shoes, and dresses are mapped to 0-9 integers. Fashion MNIST, however, has limitations in terms of real-world applications, and the DeepFashion dataset includes many additional features that enable the development of more powerful algorithms in clothing recognition. DeepFashion is a comprehensive dataset that contains over 800,000 diverse fashion images ranging from online shop images to street snapshots and consumer photos. It is also annotated with rich information of clothing items, with each image in this dataset labeled with 50 categories, 1,000 descriptive attributes, bounding box and clothing landmarks.

This project is an attempt use the above datasets in order to devise an optimal network model for fashion classification, and to also challenge ourselves to devise a real-time clothing recognition system trained on our network model. As noted in previous papers, this task, especially for real-world applications, such as video surveillance, remains challenging [2]. This is largely due to differences among various clothing categories being vague even for humans, and thus requiring considerable computations in order to discern them. As a result, this paper embodies an experimentation with various methods of creating a real-time clothing recognition system, and discusses the development choices that were made as a result of experimentation.

The ablation studies we conducted as part of our experimentation and development process include comparing the results of using different data sets (i.e. Fashion MNIST, Deep Fashion), using pre-trained weights from ImageNet vs. random initialization, using different networks, using different learning weights, and passing our trained network through autoencoders.

II. Related Work

Clothing classification and the use of clothing as a contextual cue for people identification purposes has become an increasingly popular topic of research in computer vision. Other research applications include its potential use for fashion design and e-commerce [7], whether the goal is to accurately classify clothes for e-commerce images, or or to match a real-world example of a garment to the same item in an online shop [3]. A number of different approaches have been taken in order to solve this classification problem, such as experimenting with deep learning baseline methods and learned similarity measures [3], creating a clothing segmentation method with Voronoi images to select seeds for region growing [2], and developing computer vision algorithms that describe objects by their semantic attributes [1].

Since one of our goals was to experiment with using both the Fashion MNIST and DeepFashion datasets to train

a network model, our main initial reference was a comprehensive study on the DeepFashion dataset [6]. The paper demonstrated the advantages of DeepFashion by using the dataset to train an effective model, called "FashionNet," that learns clothing features by jointly predicting clothing attributes and landmarks [6]. The estimated landmarks are then used to pool or gate the learned feature maps, leading to more robust and discriminative representations for clothing items [6]. The network structure used for the "FashionNet" model was noted to be very similar to VGG-16, which is highly effective with various vision tasks like recognition and segmentation [6]. We used this information in order to experiment with using the VGG-16 network for our own recognition system.

For the real-time video recognition feature of our program, we referenced a paper that implements a real-time clothing recognition program for surveillance videos [2]. The paper acknowledges that clothing recognition for real-time surveillance is a difficult task that involves several sub-tasks, including localization of human figures, clothing segmentation and alignment, and extraction of clothing representation [2]. It notes that due to the fact that each of the sub-tasks previously mentioned are not yet fully solved problems for surveillance videos, we cannot expect reliable results from every single frame [2]. As a result, they opt to collect clothing instances on the trajectory of a person and preserve only the top N good instances that are measured by non-occluded areas and the quality of segmentation results–the average features of those instances are then used to create clothing representations [2]. We came across similar results in terms of the accuracy of our real-time recognition system at each frame, and also made note of the differences in our accuracy rates when considering every frame vs. the top N good instances.

A potential solution to the unresolved problems discussed in the previous paper on video surveillance [2] is introduced in another paper that proposes an automated system that learns semantic attributes for clothing on the human upper body using pose estimation [1]. It takes into consideration the diversity of clothing attributes, and that a single feature is not likely to perform well on all attributes–instead, for each attribute, the prediction is obtained by compiling the classification results from several complementary features [1]. Clothing attributes are naturally correlated, which allows for the identification of mutual dependencies between attributes [1]. Although we were not able to add similar considerations to our program, we found this to be an interesting approach and solution to clothing classification.

III. Methods

A. Datasets

For testing and for our own edification, we developed network models trained on two different datasets, the Fashion MNIST dataset and the DeepFashion dataset.

- Fashion MNIST is a standard dataset of clothing article images and consists of a training set of 60,000 28x28 grayscale images and a test set of 10,000 images. Ten types of clothing, such as t-shirts, shoes, and dresses are mapped to 0-9 integers.
- DeepFashion is a comprehensive dataset that contains over 800,000 diverse fashion images ranging from online shop images to street snapshots and consumer photos. It is also annotated with rich information of clothing items, with each image in this dataset labeled with 50 categories, 1,000 descriptive attributes, bounding box and clothing landmarks.

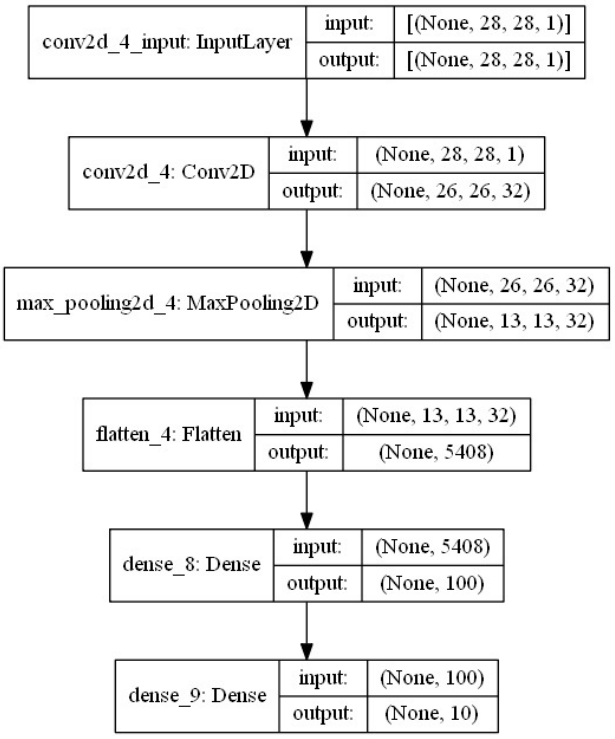

Fig. 1. CNN for training with the Fashion MNIST dataset.

B. Network Models

For developing the network model with the Fashion MNIST dataset, we used a CNN for classification with the layers listed below and as seen in Figure 1:

- Sequential
- Conv2D
- MaxPooling2D
- Flatten
- Dense (relu)
- Dense (softmax)

In order to train the network model with the DeepFashion dataset, we experimented with several models listed below:

- ResNet
- Inception ResNet
- MobilNet
- Vgg

For each of the models, we used adam optimizer with a learning rate of 0.01. All four models are pre-built networks in tensorflow, and for each of the models we used both

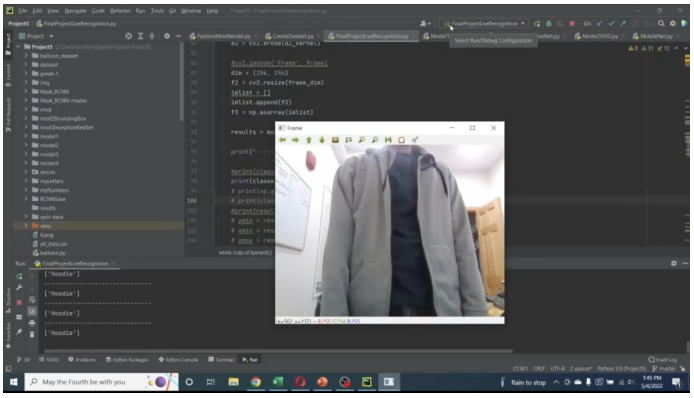

Fig. 2. Screenshot of our working real-time recognition system.

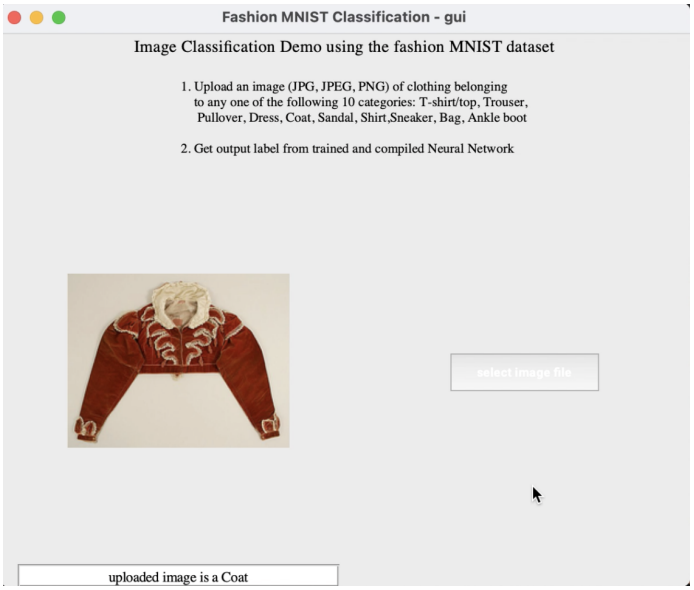

Fig. 3. GUI system displaying a correct classification label for an image of a coat referenced from the MET Museum.

pre-trained weights from ImageNet for transfer learning and random initialization. We then selected the best performing network models for implementing our real-time clothing recognition program as demonstrated in Figure 2.

C. GUI

In order to further facilitate testing different combinations of datasets and models for implementing our final real-time recognition system, we decided to develop a simple GUI that would enable us to pass various images of clothing items through different network models trained on either Fashion MNIST or DeepFashion. An example image of the GUI correctly identifying a coat garment can be seen in Figure 3.

IV. Experiments and Results

We experimented with two different datasets, Fashion MNIST and DeepFashion. The Fashion MNIST dataset has 60k training grayscale images of 28*28 size and 10k testing images while Deep Fashion has 800k diverse fashion images ranging from well-posed shop images to unconstrained consumer photos. We also used a basic CNN for the Fashion MNIST dataset to classify the data, while we trained four different models (ResNet, Inception Resnet, VGG, MobileNet) with the DeepFashion dataset. For each of these models we used adam Optimizer and after some testing arrived at 0.01 as the optimal learning rate. All four of these models are pre-built networks in tensorflow, and for each of these we experimented with using either pre-trained weights from ImageNet (transfer learning) or random initialization.

In terms of results, our models trained on the Fashion MNIST dataset performed at an accuracy of 0.9 for classification; however, the classification accuracy for the models trained on the DeepFashion dataset were less satisfactory–all of models performed at an average of 0.6 accuracy. We noted that upon looking at just the top few predictions, we were able to obtain a much more improved accuracy of around 0.8 or higher. The Inception Resnet and the Mobilenet networks ended up performing the best in terms if accuracy. Mobilenet was also much faster to train than the others. Lastly, with our experimentation with using pre-trained weights from Imagenet and without pre-trained weights or random initialization, we did not find any significant differences between using either pre-trained weights or random initialization.

V. Discussion and Summary

The final iteration of our trained deep network model obtained good results for clothing recognition. Comparing the use of our CNN with other traditional manually designed feature abstracted methods, our algorithm demonstrates much better performance. Similarly to several related papers, we acknowledge that clothing recognition for real-time surveillance is a complex task that involves several sub-tasks, including compartmentalizing human figures, clothing segmentation and alignment, and extraction of clothing representation. In future extensions of our work, we hope to optimize our deep networks architecture to improve the accuracy of real-time recognition. Examples of ways in which we can do this include factoring in human pose estimation and facial recognition as recommended by other works [2][1] in order to improve our system's ability to segment different clothing items layered onto the human body. The clothing landmarks and bounding box information provided by the DeepFashion dataset are also left-out elements that may be essential for developing a more robust classification system [6]. As briefly outlined in *Related Works*, clothing classification is an increasingly popular research topic in computer vision that proposes many useful applications from e-commerce to video surveillance. Having gained a more comprehensive introduction to clothing classification as a computer vision problem through our own experimentation and an already extensive pool of related work, we hope to be able to continue to build upon our current implementation.

REFERENCES

[1] H. Chen, A. Gallagher, and B. Gerod, "Describing clothing by semantic attributes," in *ECCV*, Springer-Verlag Berlin Heidelberg, 2012, pp. 609–623, 2012.

[2] M. Yang and K. Yu, "Real-time clothing recognition in surveillance videos," in *ICIP*, 2011, pp. 2937–2940.

[3] M.H. Kiapour, X. Han, S. Lazebnik, A.C. Berg, T.L. Berg, "Where to buy it: matching street clothing photos in online shops," in *ICCV*, Santiago, Chile, Dec 2015, pp. 3343-3351.

[4] T. Mallavarapu, L. Cranfill, E. Kim, R. Parizi, J. Morris, J. Son, "A federated approach for fine-grained classification of fashion apparel," in *Machine Learning with Applications*, 2021.

[5] D. Anguelov, K. Lee, S.B. Gokturk, B. Sumengen, "Contextual identity recognition in personal photo albums," in *CVPR*, 2007.

[6] L. Ziwei, P. Luo, S. Qui, X. Wang, X. Tang, "DeepFashion: powering robust clothes and retrieval with rich annotations," in *CVPR*, 2016.

[7] J. Cychnerski, A. Brzeski, A. Boguszewski, M. Marmolowski and M. Trojanowicz, "Clothes detection and classification using convolutional neural networks," in ETFA, 2017, pp. 1-8.