Fashion Attribute Classification Using Deep Learning

by Elizabeth Fu, Amanda Hioe, Idris Wardere, Lufei Yang
Project Supervisors: Aaditya Ramdas, Gonzalo Mena | Client: Pendulum Fashion

Introduction

We aim to enhance the accuracy of fashion trend predictions and help retailers identify fashion patterns by comparing deep learning techniques to classify dresses by length.

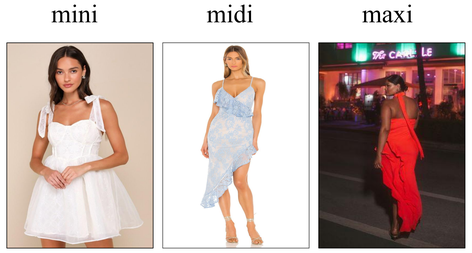
Data

The dataset employed in this study consists of 11,976 images of women's dresses, provided by our external client, Pendulum Fashion. These images encompass a diverse spectrum of dress lengths, styles, and colors, categorized into three primary dress lengths:

(1) **mini** (defined as dress hem falling above the knee)
(2) **midi** (defined as dress hem falling between the knee and ankle)
(3) **maxi** (defined as dress hem falling below the ankle)

The images in this dataset exhibit significant variation in terms of image resolution, background complexity, and the number of dresses per image. This diversity is helpful for training robust machine learning models as it introduces a realistic spectrum of scenarios that models might encounter in practical applications.

1. Examples of Diversity of Images



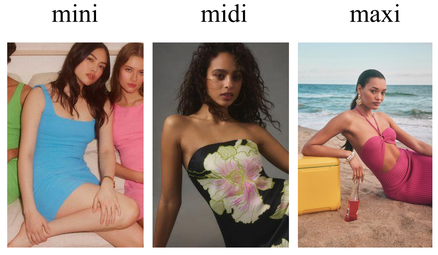
Shown in Figure 1 to the left are examples of the diversity of dress portrayal in the images in the dataset:

- (left) non full-body images
- (middle) blank backgrounds
- (right) different posing & light

The exploratory data analysis (EDA) process included assessing the distribution of dress lengths, identifying anomalies or outliers. We found each category represented by approximately 4,000 images, indicating a well-balanced dataset that aids in avoiding classification bias.

There are examples of outlier images that are hard to remove from the set. As shown in Figure 2 on the right, these images make classification difficult even for humans due to multiple models (left), difficult cropping (middle), or inclusion of other props or distracting backgrounds (right).

2. Examples of Outlier Images



Methods

We leverage four different deep learning architectures for image classification:

(1) a convolutional neural network (CNN)
(2) a residual neural network (ResNet)
(3) OpenAI's Contrasting Language-Image Pretraining (CLIP)
(4) a vision transformer (ViT)

Results

Summarized in Figure 3 to the right, a custom from-scratch CNN architecture and a fine-tuned ResNet-50 model demonstrated best performance, with overall validation accuracy of 85.7% and 86.7%, respectively. Conversely, the best CLIP and fine-tuned ViT models exhibited lower validation accuracy, at 56.0% and 56.7% respectively.

3. Validation Performance Metrics Comparison

Table 1: CNN

Class	Precision	Recall	F1-Score
mini	0.911	0.950	0.930
midi	0.836	0.805	0.820
maxi	0.820	0.816	0.818
Average	**0.856**	**0.857**	**0.856**
Accuracy		**0.857**	

Table 2: Fine-tuned ResNet50

Class	Precision	Recall	F1-Score
mini	0.890	0.954	0.921
midi	0.820	0.882	0.850
maxi	0.888	0.766	0.827
Average	**0.866**	**0.867**	**0.866**
Accuracy		**0.867**	

Table 3: CLIP Model

Class	Precision	Recall	F1-Score
mini	0.700	0.369	0.475
midi	0.206	0.053	0.084
maxi	0.407	0.908	0.562
Average	**0.438**	**0.443**	**0.374**
Accuracy		**0.560**	

Table 4: Vision Transformer

Class	Precision	Recall	F1-Score
mini	0.628	0.729	0.675
midi	0.529	0.393	0.451
maxi	0.529	0.529	0.553
Average	**0.562**	**0.550**	**0.560**
Accuracy		**0.567**	

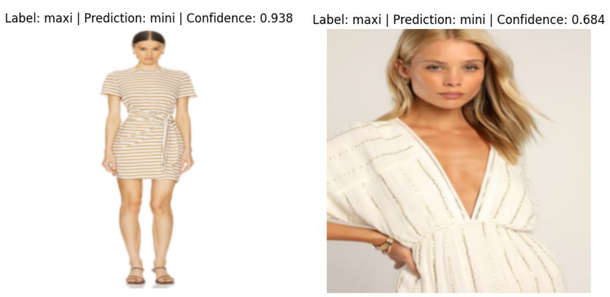
Discussion & Conclusions

We conclude that the CNN and ResNet-50 architectures both produced the similarly positive results with respect to the four different metrics used, however the ResNet-50 produced the best accuracy. For both the CNN and ResNet, the F1-score for mini dresses tended to be the highest, whereas both midi and maxi dresses were significantly worse.

Across both the CLIP and ViT models, the lower recall score, especially for midi dresses, is the major contributor to lower performance. In particular, low recall suggests generally lower rate of guessing this class. It reveals these models to tend to guess dress lengths at the extremes (mini and maxi), as opposed to in the middle (midi).

Of the better performing models, higher classification performance of mini dresses in the ResNet-50 and CNN suggest that both these models struggle with the classification of longer dresses. Many of these images were mislabeled in the original dataset (Figure 4, left) or cropped such that a human cannot correctly identify the length (Figure 4, right).

4. Examples of Incorrectly Classified Images



Many misclassified images (distribution for ResNet-50 shown in Figure 5 to right) were predicted with high confidence, suggesting the need to consider data labeling quality for further exploration and improvement on accuracy.

5. Confidence of Incorrect Predictions