
Clothing Garment Classification Using a CNN

Roy Bou Abboud

1009905462, roy.bouabboud@mail.utoronto.ca

[Link to Google Colab](#)

Project Description

The goal of my project is to create a supervised deep learning (DL) model that classifies the type of clothing garment in an image, specifically using a convolution neural network (CNN). As an Engineering Science Robotics student, the motivation for this project lies in eventually building a cloth-folding robot, capable of using a computer vision system to classify clothes and make the appropriate folds. This robot would be gifted to my mom, ensuring she never folds clothes again, thus explaining my motivation. Hence, this project is useful as a subsystem to the robot in the future, yet my scope will remain entirely focused on the DL model for garment classification as required in APS360. I believe the use of DL is appropriate for this task as DL models excel in classification problems and are widely used for such purposes, and the nature of this problem is inherently a question of classification. It is common practice to process image data in a DL model and predict the classes of objects within it using a CNN.

Illustration

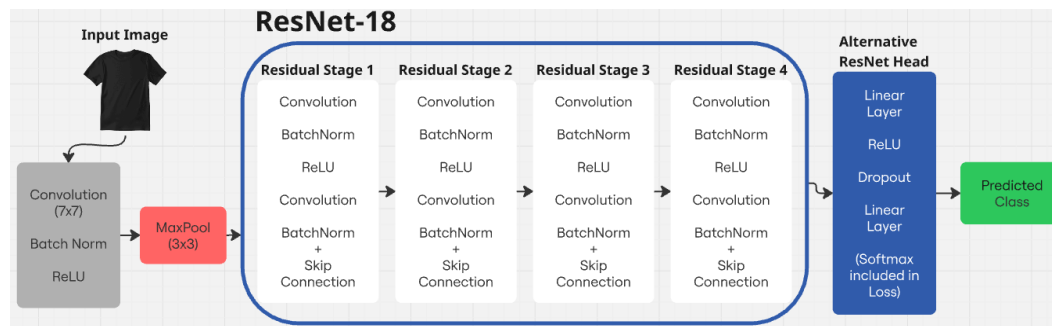


Figure 1: Flowchart of DL model workflow for garment classification.

Background & Related Work

1. **DeepFashion and FashionNet** [1]
Researchers developed a large dataset of clothing images (FashionNet) and demonstrated the effectiveness of DL in garment classification.
2. **Clothing Attribute Classification** [2]
CNNs were used to determine clothing attributes and identify similar products.
3. **HOG & SVM Method** [3]
This work showed the effectiveness of HOG + SVM prior to DL, which is a good baseline.
4. **ImageNet Classification** [4]
CNNs outperform traditional methods in large-scale image classification, leading to architectures like ResNet (used in this project).
5. **Clothing Garment Recommendations** [5]
CNNs were used to recommend garments based on images and compute compatibility scores.

Data Processing

Three datasets were used, all acquired from Kaggle. The images that were collected as data were constrained to only include one garment at a time, from a direct angle, that is not worn by any person. The six classes included t-shirts, long sleeve shirts / sweaters, hoodies, long pants, shorts, and skirts. The primary dataset [6] provided images for all classes, and the second [7] and third [8] datasets were used to supplement the data in the long pants, shorts, skirts, and hoodie classes (Figure 2). These datasets required tedious manual filtering to remove images that included people. In the end, the final classes were largely imbalanced as seen in Figure 2 (1874 long pants VS 227 skirts), which was overcome by hard capping large classes and augmenting small classes to 1000. By mixing many datasets, I created a diverse yet unified set of images that fit the constraints, and that can further be split into training / validation / testing sets.

Garment Type	Dataset 1	Dataset 2	Dataset 3	Total (Combined)
long_sleeve_sweater	1077	0	0	1077
tshirt	1011	0	0	1011
long_pants	692	1182	0	1874
shorts	308	521	0	829
skirts	155	72	0	227
hoodie	100	436	527	1063

Figure 2: Total number of data points (images) in each class based on combined dataset

Primary Model Architecture

As seen in the illustration (Figure 1), the primary model for this project is a CNN built using a transfer learning ResNet-18 model with a custom fully connected linear head. This model boasts a total of 11,309,894 parameters, 8,527,110 of which are trainable. After tuning, the optimized model used a linear layer hidden dimension of 128, dropout of 0.4, learning rate of 0.001, weight decay of $1e-5$, and batch size of 16. The full architecture including the custom head can be seen in Figure 1.

Baseline Model

The baseline architecture used in this project consists of a non-DL based HOG + SVM approach. HOG takes the images and converts them into a long vector of numbers that describes the features of an image, hence capturing the outlines of the clothing garments. These vector were then inputted into an SVM, which separates the data into different classes and then predicts the type of garment in the image. This setup required no tuning at all.

Quantitative Results

See Figures 3 & 4. The analysis of the quantitative results can be found in the Discussion section.

	precision	recall	f1-score	support		precision	recall	f1-score	support
hoodie	0.9388	0.9517	0.9452	145	hoodie	1.0000	0.9600	0.9796	150
long_pants	0.9205	0.9267	0.9236	150	long_pants	0.9286	0.9533	0.9408	150
long_sleeve_sweater	0.9231	0.9600	0.9412	150	long_sleeve_sweater	0.8868	0.9400	0.9126	150
shorts	0.9020	0.9200	0.9109	150	shorts	0.9247	0.9000	0.9122	150
skirts	0.8333	0.6667	0.7407	30	skirts	0.9552	0.8533	0.9014	150
tshirt	0.9514	0.9133	0.9320	150	tshirt	0.8834	0.9600	0.9201	150
accuracy	0.9239			775	accuracy	0.9278			1
macro avg	0.9115	0.8897	0.8989	775	macro avg	0.9298	0.9278	0.9278	900
weighted avg	0.9234	0.9239	0.9231	775	weighted avg	0.9298	0.9278	0.9278	900

Figure 3: Classification performance of HOG + SVM (left) VS CNN (right) on testing data

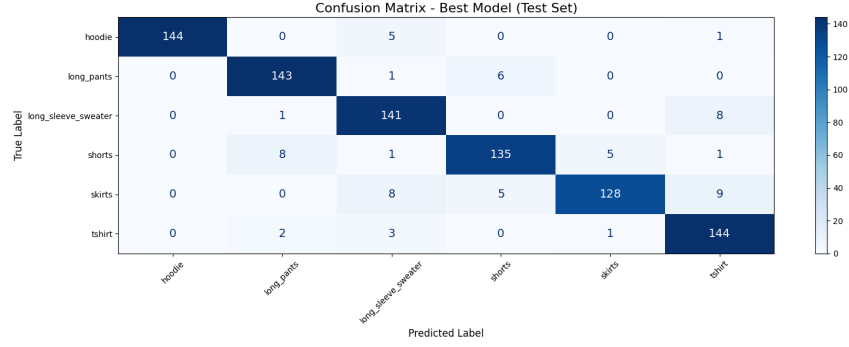


Figure 4: Confusion Matrix of CNN model

Qualitative Results

As seen from Figure 5, the model is capable of classifying diverse and complex garments. It correctly identifies the subtle features which distinguish hoodies from long sleeves, performed well for the smaller class of skirts, and is flexible enough to return a result even when a person is present. Conversely, the model struggled with ambiguous items such as short sleeve hoodies, garments that had unclear features like the sweater, and misinterpreted smaller shorts as skirts.



Figure 5: Correctly (top) and incorrectly (bottom) predicted garments

Evaluation of Model on New Data

Within this project, a test split with unseen data was used as a final check to determine the model's accuracy. My best model returned a test loss of 0.2396 and test accuracy of 0.9278, which indicates strong performance yet similar to the baseline model. Beyond that, as mentioned in my progress report, a true test of utility for my model would be to test it using clothing garments found in my closet, as these items will be repeatedly folded by a future robot. Hence, I took three different pictures of a garment in my closet from each class and attempted to use my CNN to classify them. These images followed all the previous constraints that were used in data processing. Fortunately, the CNN was able to correctly class all 18 pictures with extremely high confidence as seen in Figure 6 and Appendix A. This demonstrates that on real-world data that is relevant and never seen before, the model shows signs of good performance. However, 18 images is still a small sample and this is not enough evidence to claim that the system is generally robust for real-world applications.



Figure 6: Model predictions VS ground truths of garments from my (and sister's) closet

Discussion

The results above suggest that my model is performing well due to multiple factors. As a first check, the CNN earned a test accuracy of 0.9278, which is minimally larger than the HOG + SVM baseline of 0.9239. More importantly, my model earned a macro-average F1 score of 0.9278 vs 0.8989 for the baseline model, meaning the CNN generalizes better across all classes. This is the true advantage of my model: individual class performances. The F1 score for skirts improved from 0.7407 to 0.9014, which is significant due to the ambiguity of the garment in comparison to others and its small class size. An incredible result was the F1 score of 0.9796 for hoodies using the CNN, which further validates the qualitative claim that my model can successfully learn the necessary features to distinguish between hoodies and long sleeves / sweaters. These improvements did come with a small decrease in the F1 scores for t-shirts (0.9320 vs 0.9201) and long sleeves / sweaters (0.9412 vs 0.9126). Hence, the most significant inference to be made is that the CNN trades small declines in strong performing classes for large improvements in poor performing classes. This is a crucial distinction, as a cloth folding robot should generalize well for all individual classes instead of being perfect at some and bad at others. As seen from the confusion matrix, the highest concentration of truth-prediction pairs are aligned in the diagonal, which indicates good performance. The weakest diagonal was the skirts with 128/150 correct predictions, which is expected since it is the most ambiguous class. The errors found in the confusion matrix are not random, but rather explainable, like confusing long pants as shorts due to the angle of the pictures or length of the garment, and vice-versa. This supports the qualitative claim that unclear features may corrupt the model's prediction.

Ethical Considerations

A major ethical concern is bias towards western clothing styles, as the data used in this project over-represent these garments. Hence, it could not be considered a universal garment classifier. This concern becomes much graver if this CNN classifies people's garments in the real world. Diversifying data collection and including more classes would be a reasonable strategy to mitigate this concern.

Project Difficulty

At face value, this project may seem simple, however it was very nuanced and complex. Data processing was challenging as images had to be tediously filtered by hand. Distinguishing between classes such as hoodies & sweaters or shorts & skirts required a strong model that was fine-tuned to detect specific edges and garment shapes. Finally, creating a HOG + SVM baseline model from scratch and editing the ResNet's head required a combination of course content and extracurricular research, exploring concepts beyond the scope of this course.

References

- [1] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 1096–1104. doi: 10.1109/CVPR.2016.124.
- [2] J. Huang, R. Feris, Q. Chen, and S. Yan, “Cross-Domain Image Retrieval with a Dual Attribute-Aware Ranking Network,” in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile: IEEE, Dec. 2015, pp. 1062–1070. doi: 10.1109/ICCV.2015.127.
- [3] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), San Diego, CA, USA: IEEE, 2005, pp. 886–893. doi: 10.1109/CVPR.2005.177.
- [4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2012. Accessed: Jan. 26, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html
- [5] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, “Learning Fashion Compatibility with Bidirectional LSTMs,” in Proceedings of the 25th ACM international conference on Multimedia, Oct. 2017, pp. 1078–1086. doi: 10.1145/3123266.3123394.
- [6] “Clothing dataset (full, high resolution).” Accessed: Apr. 07, 2026. [Online]. Available: <https://www.kaggle.com/datasets/agrigorev/clothing-dataset-full>
- [7] “Clothing & Models.” Accessed: Mar. 13, 2026. [Online]. Available: <https://www.kaggle.com/datasets/dqmonn/zalando-store-crawl>
- [8] “Apparel Dataset.” Accessed: Mar. 13, 2026. [Online]. Available: <https://www.kaggle.com/datasets/kaiska/apparel-dataset>

Appendix A

Performance of the CNN model on the unseen closet test set:

Closet Test Results (Accuracy: 100.0%)					
	File	True Label	Predicted	Confidence	Correct
0	hoodie_1.jpg	hoodie	hoodie	86.49%	✓
1	hoodie_2.jpg	hoodie	hoodie	99.60%	✓
2	hoodie_3.jpg	hoodie	hoodie	82.49%	✓
3	long_sleeve_sweater_1.jpg	long_sleeve_sweater	long_sleeve_sweater	99.84%	✓
4	long_sleeve_sweater_2.jpg	long_sleeve_sweater	long_sleeve_sweater	96.10%	✓
5	long_sleeve_sweater_3.jpg	long_sleeve_sweater	long_sleeve_sweater	99.68%	✓
6	pants_1.jpg	long_pants	long_pants	99.92%	✓
7	pants_2.jpg	long_pants	long_pants	83.14%	✓
8	pants_3.jpg	long_pants	long_pants	99.91%	✓
9	shorts_1.jpg	shorts	shorts	99.79%	✓
10	shorts_2.jpg	shorts	shorts	99.17%	✓
11	shorts_3.jpg	shorts	shorts	99.16%	✓
12	skirt_1.jpg	skirts	skirts	99.95%	✓
13	skirt_2.jpg	skirts	skirts	99.94%	✓
14	skirt_3.jpg	skirts	skirts	99.39%	✓
15	tshirt_1.jpg	tshirt	tshirt	99.86%	✓
16	tshirt_2.jpg	tshirt	tshirt	99.86%	✓
17	tshirt_3.jpg	tshirt	tshirt	99.68%	✓