COSE474-2024F: Final Project Report
Fashion Item Classification with ResNet and CLIP

Aisyah
2022320119

1. Introduction

Fashion item classification is a critical task in e-commerce, enabling applications such as automated product categorization, recommendation systems, and personalized shopping experiences. Traditional deep learning models like ResNet18 require large labeled datasets for training, which may not be available in many real-world scenarios. Vision-language models like CLIP have shown promise for zero-shot learning by leveraging large-scale pretraining on image-text pairs. These models have demonstrated robust performance in diverse tasks, making them an ideal candidate for fashion classification.

1.1. Motivation

Automating fashion classification can:

1. Enable personalized outfit recommendations.
2. Enhance operational efficiency for online stores by automating inventory categorization.
3. Improve user experience by providing quick and accurate fashion suggestions.

This project evaluates CLIP and ResNet18, focusing on the effectiveness of CLIP's pre-trained embeddings in data-scarce scenarios, inspired by the findings in which highlight the utility of pre-trained models in vision-language tasks.

1.2. Problem Definition

The goal is to classify images into three categories: shirts, dresses, and shoes. Challenges include:

1. **Limited labeled data:** The small dataset makes supervised learning difficult.
2. **Visual ambiguity:** Categories like shirts and dresses share visual similarities, complicating classification.
3. **Generalization:** Ensuring that models perform well on unseen data.

1.3. Contribution

This report evaluates two approaches:

1. Fine-tuning ResNet18 for supervised classification.
2. Using CLIP for zero-shot classification with image-text prompts.

The primary contribution is to compare the performance of both methods, which discusses the role of distribution-aware prompts in enhancing model generalization.

2. Methods

2.1. Significance and Novelty

This project combines traditional supervised learning (ResNet18) with a modern approach (CLIP), which leverages pre-trained vision-language embeddings. CLIP's ability to align vision and language representations is particularly advantageous for zero-shot learning. This dual-model comparison provides insights into how pre-trained embeddings can outperform supervised models in scenarios with limited labeled data.

2.2. Implementation Details

ResNet18: Fine-tuned for 5 epochs using the Adam optimizer and cross-entropy loss. Data augmentation techniques like rotation, flipping, and color jittering were used to combat overfitting.

CLIP: The pre-trained CLIP model was used in a zero-shot setting, utilizing text prompts such as "A photo of a shirt." The effectiveness of prompts was tested without additional fine-tuning, relying solely on pre-trained embeddings.

3. Experiments

3.1. Dataset

The dataset consists of 12 training images and 4 testing images for each category (shirts, dresses, shoes). Data augmentation techniques such as random rotations, horizontal flipping, and color jittering were applied to artificially expand the training set, addressing the small dataset limitation.

3.2. Computing Resources

1. **Hardware:** Google Colab with Tesla T4 GPU.
2. **Software:** Python 3.10, PyTorch 2.0, torchvision 0.15, transformers 4.32.
3. **OS:** Ubuntu 20.04.

3.3. Quantitative Results

Table 1 summarizes the performance of ResNet18 and CLIP on the test dataset. CLIP's superior accuracy reflects its ability to generalize effectively.

Table 1. Performance Comparison

Model	**Accuracy**	**F1-Score**
ResNet18	45.83%	0.39
CLIP	100%	1.00

3.4. Qualitative Results

Visualization of predictions highlights the strengths and weaknesses of each model. Figure 1 shows ResNet18's frequent misclassifications between "shirts" and "dresses," consistent with its struggles to handle ambiguous features.

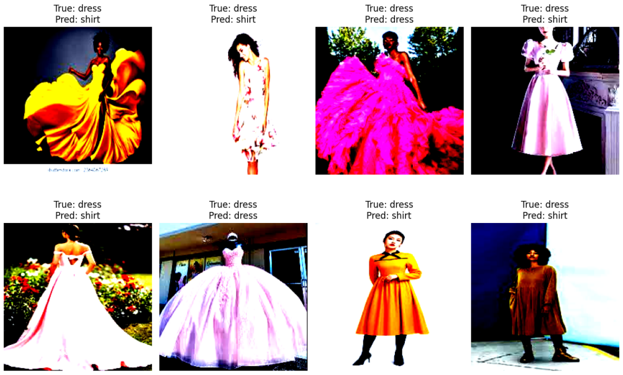

Figure 1. Sample predictions: ResNet18 misclassified "shirts" and "dresses."

3.5. Confusion Matrix

The confusion matrix in Figure 2 illustrates ResNet18's performance. It reveals that most errors occurred in misclassifying "dresses" as "shirts," a challenge CLIP was able to overcome.

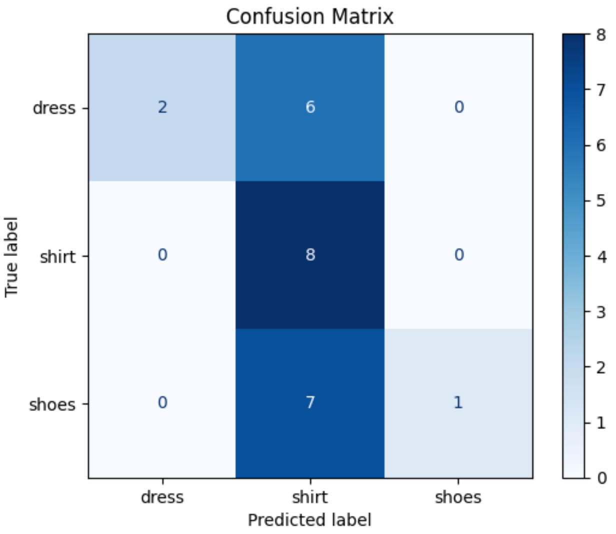

Figure 2. Confusion matrix for ResNet18.

3.6. Discussion

ResNet18 Observations:

1. ResNet18 struggled with limited labeled data, resulting in significant misclassifications.
2. Data augmentation slightly improved generalization but could not overcome the dataset's small size.

CLIP Observations:

1. CLIP achieved perfect accuracy, demonstrating the potential of vision-language embeddings for zero-shot classification.
2. The alignment of image and text embeddings enabled robust generalization, even on a small dataset.
3. These results align which highlights CLIP's robustness in diverse settings.

4. Future Directions

Future work should focus on:

1. Increasing the dataset size to improve ResNet18's ability to generalize.
2. Experimenting with advanced prompt engineering to enhance CLIP's performance on ambiguous categories.
3. Developing hybrid models that combine ResNet18's supervised learning with CLIP's zero-shot capabilities.

4. Fine-tuning CLIP on fashion-specific datasets to achieve even higher accuracy.

References

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., and others. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML).*

Zhang, Q., Wu, T., Chen, Y., and Luo, H. (2023). Distribution-Aware Prompt Tuning for Vision-Language Models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

Liu, H., Feng, Z., and Du, X. (2021). Pretrain and Predict: Multi-task Fine-tuning for Vision-Language Tasks. *arXiv preprint arXiv:2103.06561.*

References