Multi-Label Classification for Fashion Data: Zero-Shot Classifiers via Few-Shot Learning on Large Language Models

Dongming Jiang, Abhishek Shah, Stanley Yeung, Jessica Zhu, Karan Singh and George Goldenberg
CaaStle Inc, U.S.A.
{*dj, abhishek.shah, stanley.yeung, jessica.zhu, karan, golden*}*@caastle.com*

Keywords: Large Language Model, Few-Shot Learning, Zero-Shot Learning, Inference, Knowledge Generation, Multi-Label Classification, Scalability, Fashion Dynamics.

Abstract: Multi-Label classification is essential in the fashion industry due to the complexity of fashion items, which often have multiple attributes such as style, material, and occasion. Traditional machine-learning approaches face challenges like data imbalance, high dimensionality, and the constant emergence of new styles and labels. To address these issues, we propose a novel approach that leverages Large Language Models (LLMs) by integrating few-shot and zero-shot learning. Our methodology utilizes LLMs to perform few-shot learning on a small, labeled dataset, generating precise descriptions of new fashion classes. These descriptions guide the zero-shot learning process, allowing for the classification of new items and categories with minimal labeled data. We demonstrate this approach using OpenAI's GPT-4, a state-of-the-art LLM. Experiments on a dataset from CaaStle Inc., containing 2,480 unique styles with multiple labels, show significant improvements in classification performance. Few-shot learning enhances the quality of zero-shot classifiers, leading to superior results. GPT-4's multi-modal capabilities further improve the system's effectiveness. Our approach provides a scalable, flexible, and accurate solution for fashion classification, adapting to dynamic trends with minimal data requirements, thereby improving operational efficiency and customer experience. Additionally, this method is highly generalizable and can be applied beyond the fashion industry.

1 INTRODUCTION

Multi-label classification plays a crucial role in fashion applications due to the complex nature of fashion items, which often possess multiple attributes such as style, material, occasion, and season. For example, a single dress might be labeled as "casual," "floral," "cotton," and "summer." Accurate classification is fundamental for various functions, including merchandising, inventory management, trend analysis, and personalized customer experiences. An efficient multi-label classification system can significantly enhance operational efficiency, customer satisfaction, and sales by aligning products with consumer preferences.

This paper presents our work in addressing multi-label classification for CaaStle Inc., a company that provides advanced technology and services to apparel brands, focusing on optimizing business operations and consumer engagement. CaaStle manages a vast inventory where garments often carry multiple labels, with some labels being far less frequent than others. The imbalanced, high-dimensional, and sparse nature of this data creates challenges for traditional machine learning approaches. Moreover, with new styles and items constantly entering the inventory, the need for continuous re-labeling and model retraining becomes costly and time-consuming.

To address these challenges, we propose a novel approach that utilizes the reasoning capabilities of Large Language Models (LLMs) to enhance multi-label classification. By integrating few-shot and zero-shot learning, our system can effectively classify new and existing fashion items with minimal labeled data. We demonstrate this approach using OpenAI's GPT-4 on a real-world dataset from CaaStle, showcasing improved classification performance and scalability. This solution adapts to fashion trends with minimal data requirements and offers potential applications beyond the fashion industry.

To our knowledge, no prior work has combined the three elements of LLMs, few-shot learning, and zero-shot learning for multi-label classification in the fashion industry. This novel integration marks a significant advancement in the field. Specifically, we are the first to leverage LLMs to generate detailed and

Jiang, D., Shah, A., Yeung, S., Zhu, J., Singh, K. and Goldenberg, G.
Multi-Label Classification for Fashion Data: Zero-Shot Classifiers via Few-Shot Learning on Large Language Models.
DOI: 10.5220/0012899100003838

In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024) - Volume 1: KDIR*, pages 250-257
ISBN: 978-989-758-716-0; ISSN: 2184-3228

precise descriptions of new fashion categories using few-shot learning. These descriptions serve as guidelines for zero-shot learning, enabling accurate classification of emerging categories.

2 RELATED WORK

The fashion industry has seen significant growth and evolution in classification techniques over the past few decades (Abbas et al., 2024; Saranya and Geetha, 2022; Abd Alaziz et al., 2023; Xhaferra et al., 2022; Guo et al., 2019a; Kolisnik et al., 2021; Q. Ferreira et al., 2019; Inoue et al., 2017; Ferreira et al., 2021). Traditional classification techniques in the fashion industry primarily relied on manual categorization, for example, based on silhouette and shapes that characterize a garment's outlines and fit, garment types and purposes such as top, dress, and pants, and design elements as well as detailed attributes of a garment style such as hemline length and neckline shape. Moving into the 21st century, the fashion industry began to adopt more sophisticated hierarchical taxonomies and categorization systems to organize garments into multiple levels using various semantic grouping and logic. Recent research has focused on hierarchical multi-label classification models (Seo and Shin, 2019; Zhong et al., 2023; Mallavarapu et al., 2021; Al-Rawi and Beel, 2020) that mimic human classification processes, and predict and produce multiple labels at different taxonomy levels for each garment. With the advent of computer vision and deep learning, more advanced and automated classification approaches like Convolutional Neural Networks (CNNs) (LeCun et al., 1998; Krizhevsky et al., 2017; Szegedy et al., 2015; He et al., 2016) have emerged, enabling image-based classification of garments, styles, and attributes directly from visual data. More recently, inspired by the rapid advancement and widespread adoption of Artificial Intelligence (AI) foundation models, application of the multi-modal techniques (Guo et al., 2019b; Ngiam et al., 2011; Lu et al., 2019) with the ability to understand and generate data across multiple modalities, for example, text and image, has become active research in fashion classification.

However, due to the complexity of algorithms that require vast amounts of training data and substantial computational power, current techniques face significant challenges in addressing the rapidly evolving dynamics of the fashion industry, particularly in classification problems. In this paper, we introduce a novel approach to multi-label classification, integrating LLMs (Chen et al., 2020), few-shot learning (Kadam and Vaidya, 2020), and zero-shot learning (Raffel et al., 2020) to develop a scalable, accurate, and flexible system tailored to the dynamic, trend-sensitive nature of fashion.

3 APPROACH

We describe our algorithm and demonstrate an implementation in more detail in this section.

3.1 Algorithm

3.1.1 Step 1: Leveraging LLM for Few-Shot Learning

1. Initial training with few-shot learning
   - Utilize a small, labeled dataset to train the LLM on specific fashion categories.
   - The LLM learns from this limited data to understand and identify key attributes and features associated with each category.
2. Inference and reasoning
   - The LLM applies its inference and reasoning capabilities to generalize from the few examples provided.
   - It identifies patterns, trends, and unique characteristics of the fashion items within the limited data, improving its understanding of the categories.

3.1.2 Step 2: Generating Descriptions for New Classes

1. Guiding LLM to generate descriptions
   - When a new fashion category and class is introduced, the LLM uses its learned knowledge and the few-shot learning context to generate a detailed and precise description of the new class.
   - This description includes key attributes, styles, materials, and other relevant features that define the new category.
2. Semantic enrichment
   - The generated description can be enriched with semantic information, leveraging embeddings and attributes that the LLM has learned from existing data.

3.1.3 Step 3: Zero-Shot Learning with Generated Descriptions

1. Utilizing descriptions for zero-shot learning

- The detailed class description generated by the LLM serves as a guideline for the zero-shot learning process.
- The system uses the description to map features of unseen instances to the new class, leveraging semantic similarities and relationships.

2. Building binary classifiers
   - For each new class, the system constructs binary classifiers using the LLM. These classifiers determine whether an instance belongs to the new class based on the description and semantic guidance.
   - The binary classifiers are integrated into the overall multi-label classification framework, enabling the system to handle multiple labels simultaneously.

3.1.4 Step 4: Multi-Label Classification

1. Integrating classifiers
   - The binary classifiers for new classes are combined with existing classifiers to create a comprehensive multi-label classification system.
   - The system evaluates each fashion item against all relevant classifiers to assign the appropriate labels.
2. Inference and prediction
   - During inference, the system processes new fashion items, applying both the few-shot learned models and the zero-shot classifiers guided by the LLM-generated descriptions.
   - The LLM's reasoning capabilities ensure accurate and context-aware predictions, even for classes with minimal or no labeled examples.

3.2 Implementation

There exist various options for LLMs in an implementation of our proposed approach. In this paper, we present experiments and results from one of our implementations using OpenAI GPT-4 (Achiam et al., 2023). GPT-4 is a state-of-the-art LLM that is pretrained. In addition to its proficiency in language understanding and generation, it excels in understanding context, following guidelines and instructions, logical inference, and basic reasoning.

Figure 1 shows our implementation of the approach for Step 1. Garment Info contains examples of the garments that belong to and that do not belong to the new class, in the form of the image and text descriptions of the garments. Class Info contains classification guidelines for the class, which can be

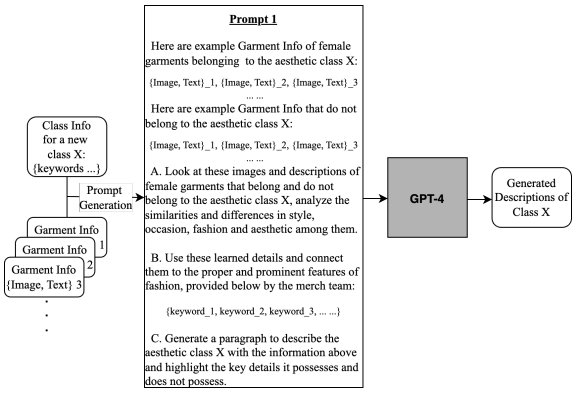

Figure 1: Implementation of the Algorithm Step 1 for running a few-shot learning with GPT-4.

in various forms that can be as simple as keywords that best describe the fashion class. Class Info and Garment Info are the inputs to GPT-4 for the few-shot learning. They can either be provided by humans or be generated by LLMs. We will compare and discuss these two different methods in more detail in the Experiment section. These inputs are structured into Prompt 1 which is sent into GPT-4 through the GPT API. The goal of Prompt 1 is to guide GPT-4 to do the few-shot learning using the labeled data and produce the class descriptions accordingly. This learning process can iterate with various examples and guidelines in multiple rounds, each of which results in a class description.

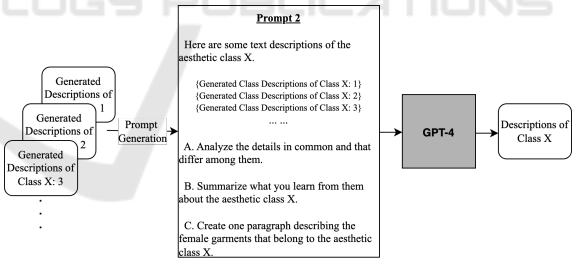

Figure 2: Implementation of the Algorithm Step 2 for generating the final descriptions of a new fashion class through GPT-4.

Figure 2 illustrates how the final descriptions of a new fashion class are generated. With potentially multiple class descriptions generated by the few-shot learning process, Prompt 2 carries these results to GPT-4. The goal of Prompt 2 is to teach GPT-4 with the knowledge that is learned from the small number of labeled data in Step 1, and instruct GPT-4 to analyze and refine them using its inference and reasoning capabilities, producing precise final class descriptions at the end.

Figure 3 demonstrates a zero-shot binary classifier. It uses Prompt 3 to instruct GPT-4 to do proper

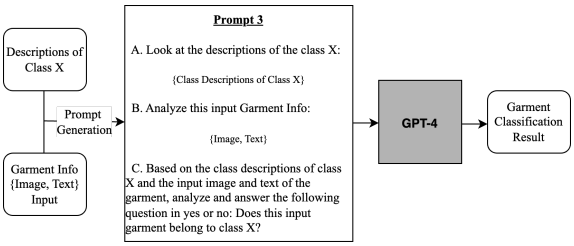

Figure 3: Implementation of the Algorithm Step 3 that builds a zero-shot binary classifier using the generated class descriptions on GPT-4.

inference and answer a binary classification question, taking into account the knowledge learned for the new fashion class and the query garment.

4 EXPERIMENT

4.1 Dataset

To support the development of the classification models and system, CaaStle picked a small proportion of its inventory pool and manually tagged and validated all of their labels. This dataset includes 2480 different styles and each style can have 1 or more of 18 different labels (Table 1). Each style comes with a vendor description and key characteristics edited by CaaStle's merchandising team. Data formats of each style include a primary image, multiple images of the same style in various views such as front view, side view, and back view, and descriptions in text. Examples of the data can be browsed at https://closet.gwynniebee.com/ and https://www.haverdash.com/. In the rest of the paper, when we refer to an image of a style, it is always the primary image. When multiple views of a style are used in certain approaches, we will explicitly call them out as multi-view images. We will refer to the edited vendor description of each style Human product description in this paper. The merchandising team also provides a natural-language description of each class / label and classification guidelines, and uses them to train the team for the manual tagging and validation of the class labels. We will call this data Human classification guidelines in this paper. Each style can be tagged with multiple classes or labels in Aesthetic Styles as well as in Occasions, and only a single class or label in Weather. In the rest of the paper, we use the terms class and label interchangeably.

Table 1: Category and Class labels in the dataset.

Aesthetic Styles	Occasion	Weather
Feminine	Party	Cold
Classic	Casual/Lounge	Warm
Edgy	Resort	Year-round
Boho	Day Night	
Retro	Work	
Athleisuren	Everyday	
Minimalist	Wedding Guest	
Preppy		

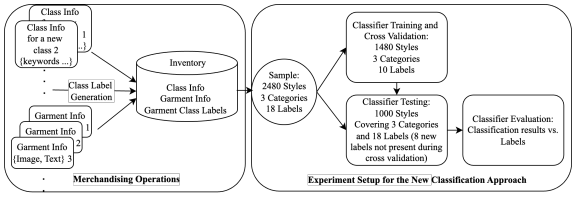

Figure 4: Workflow and setup of the experiment.

4.2 Experiment Setup

The data selection process is guided by general fashion classification criteria and the high-level distribution of style attributes, such as product types (e.g., tops, dresses, pants), fabric, labor costs, and the constraints of manual tagging and validation. The merchandising team continuously provides subsets of the dataset through the data pipeline. This approach aligns with our model exploration, testing, and system development processes. The workflow and experimental setup are illustrated in Figure 4. We use 60% of the dataset, which arrived earlier in the pipeline, for experimentation, model training, and validation. The remaining 40%, including new labels absent during the training phase, is used to test the classification approach. This setup simulates a real-world scenario where not only new styles of existing labels emerge, but entirely new classes and labels also appear over time. The classification system adapts by learning and building new classifiers for these emerging classes and labels, using a few example labels generated by the merchandising team throughout the process.

4.3 Metrics

To evaluate the classification performance, we consider three relevant metrics.

4.3.1 Accuracy

Accuracy = (True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives)

Accuracy gives a straightforward measure of overall performance. However, it can be misleading in the case of imbalanced datasets where the majority class dominates the metric.

4.3.2 F1-Score

F1-score is the harmonic mean of precision and recall. F1-score helps alleviate the bias of Accuracy towards dominant classes in imbalanced data. It is more informative than Accuracy especially when the dataset has uneven class distribution by balancing both precision and recall.

- *F1-score = 2 * (Precision * Recall) / (Precision + Recall)*
- *Precision = (True Positives) / (True Positives + False Positives)*
- *Recall = (True Positives) / (True Positives + False Negatives)*

4.3.3 Weighted F1-Score

It is insufficient to compute only the F1-score for each class independently because CaaStle judges the quality of the multi-label classification at the category level across all its classes in addition to the quality of each class. When evaluating quality, the business regards every instance of a single labeling equally, and every label equally. Therefore, we compute a weighted F1-score using a weight that reflects the proportion of the true instances from each class over the total instances of the category.

$$Weighted\ F1 = \sum_{i=1}^{N} w_i F1_i \tag{1}$$

This method takes class imbalance into account, where N is the number of classes in the category, w_i is the ratio of the number of true instances for each class to the total instances for the category, and $F1_i$ is the F1-score for each class.

We present the results in F1-scores for each class and Weighted F1-scores for each category and dataset in this paper.

CaaStle's quality target of the classification system is to achieve at least 0.7 of F1-score for each class, and 0.8 of weighted F1-score for the category that includes the classes.

4.4 Experiments on State-of-the-Art Models

With the labeled styles and their image and text data, we attempt to train a multi-label classification model, using the 2480 unique styles, text description for each of them, 10K multi-view images for all the styles, and 18 possible labels, through the typical training, validation, and testing process. This is an important task in our experiments because we need to understand whether the state-of-the-art modeling methods can support the multi-label classification requirements, and if they do not, what problems we need to address in designing the new methods. The modeling methods we test include Google Vertex by training a classifier from scratch, and three widely adopted pre-trained image classification models, ResNet-50 (Koonce and Koonce, 2021), Vision Transformer (ViT) (Dosovitskiy et al., 2020), and Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021). We use only image data for Vertex, RestNet-50, and ViT, but {image, text caption} data for CLIP to take advantage of CLIP's multi-modal capability. Experiments show common evidence of serious overfitting across all these different methods. The class-level F1-scores spread from 0.1 to 0.8, and the category-level weighted F1-scores are usually around 0.5 and below. The main challenge comes from the lack of labeled data for the multi-label classification problem. For example, during fine-tuning of the pre-trained models, we need to tune the last layer by having the number of nodes match the number of labels, using the sigmoid rather than the softmax activation function for each node, and fitting with the binary cross-entropy loss function. This more complex mathematical form of the models requires much more labeled data for training. To validate the hypothesis about the impacts of the problem complexity, we also test by reducing the complexity of the problem from multi-label to multi-class and eventually to one-vs-all classification problems. Notice that by reducing the problem complexity we also change the goal of the classification problem itself. We only do so to get a better understanding of the possible causes of the overfitting problem. Transforming the multi-label problem to a multi-class and one-vs-all classification setup indeed helps in improving the testing F1-scores, however, the overfitting is still present, and the F1-scores are still nowhere close to CaaStle's quality target. To continue in this technical direction, even for fine-tuning a pre-trained model, we will need to label a lot more styles especially styles that have multiple labels to start with. In contrast, we will show the results of our proposed approach which significantly outperforms.

4.5 Evaluating CaaStle Approaches

In this section, we summarize the key experiments and results that show the superior performance of the

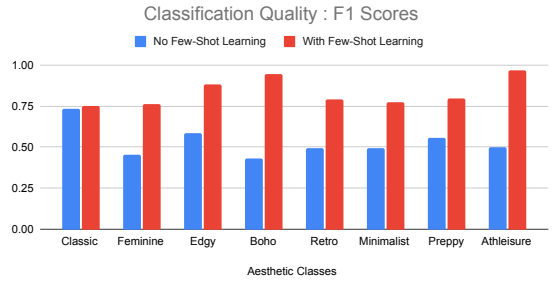

Figure 5: We compare the quality of zero-shot classification between the approaches with and without few-shot learning.

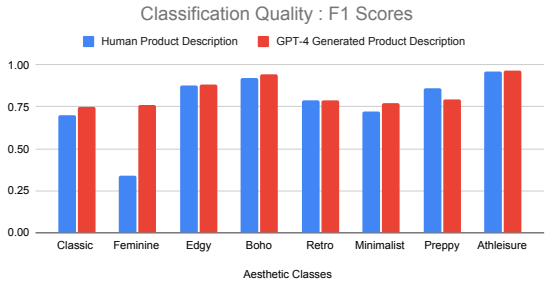

Figure 6: We compare the quality of two different methods in producing the product description of a fashion item. The product description is an important parameter for Garment Info that is required by Prompt 1 in Figure 1.

classification that is driven by integrating few-shot learning, LLM, and zero-shot learning.

4.5.1 Few-Shot Learning on LLMs Boosts Zero-Shot Classification

The crucial difference between the two zero-shot classification approaches, shown in Figure 5, is in class description generation. In the approach with no few-shot learning, we use the human classification guidelines that are crafted by the merchandising team. This approach is considered the best effort in zero-shot learning because it leverages the knowledge best known by humans. On the other hand, in the approach with few-shot learning, the classification guideline uses the class description generated by few-shot learning on GPT-4 (Figure 1, Figure 2). We are essentially comparing zero-shot binary classifiers using knowledge learned by few-shot on GPT-4 with that using human knowledge and the best efforts. The improvement in classification quality by the few-shot learning on GPT-4 is significant. Figure 5 shows that the few-shot learning always outperforms, from 2% to 118% better than the zero-shot approach without it. Even though the Aesthetic Classes are very diverse, our proposed approach of few-shot learning is quite robust, showing consistently high performance across all the classes. Compared to the other Aesthetic Classes, styles in the Classic class appear more consistent, as their characteristics are well-captured by human knowledge and descriptions. As a result, learning from additional examples does not provide significant value.

During the experimentation and related sensitivity analyses, we gain more insights into how few-shot learning and GPT-4 interplay. LLMs, including GPT-4, work well in discovering and generalizing common patterns from examples. Fashion items, however, often require attention to some subtle and seemingly minor details that can be decisive in fashion classification but not so much in machine learning. Therefore the prompt needs to be designed and experimented with to better guide GPT-4 to perform learning more specifically. The learning outcome from GPT-4 can be sensitive to the input examples. We test with various strategies, including using positive examples, negative examples, and sampled examples according to certain distribution considerations. We find that it is beneficial for running few-shot learning in multiple epochs, which allows us to run representative but diverse examples throughout the entire learning. Thereafter, we can apply different strategies and algorithms in generating the final class descriptions based on multiple candidates of the class descriptions, coming out of the few-shot learning epochs. Figure 1 and Figure 2 illustrate the prompts we design in both steps for guiding GPT-4 to perform the learning and class description generation tasks.

4.5.2 LLM Generated Garment Data Improves Classification

In the last section, we have already shown that the class description generated by LLM (GPT-4) through few-shot learning significantly improves the classification performance. In this section, we show that the classification performance is further improved by leveraging the product description that is generated by LLM (GPT-4). Figure 6 illustrates that, compared to the approach of using the product descriptions that are provided by the vendors or crafted by humans, the approach of using the GPT-4 generated texts is consistently better. The Preppy class is an exception, as the GPT-4-generated product descriptions are sometimes overly specific about certain details, which can negatively affect the class description generation. At the category level for Aesthetic Styles across all classes, using the Human Product Description yields a weighted F1-score of 0.66, while the GPT-4-Generated Product Description achieves a weighted F1-score of 0.80. This represents a 20% improvement in classification performance when us-

ing GPT-4-generated descriptions. It highlights a significant advantage of LLMs like GPT-4, which are trained on vast amounts of internet data, enabling them to reason with richer and broader contexts than the domain-specific expertise of humans.

4.5.3 Multi-Modality Improves Few-Shot Learning Performance

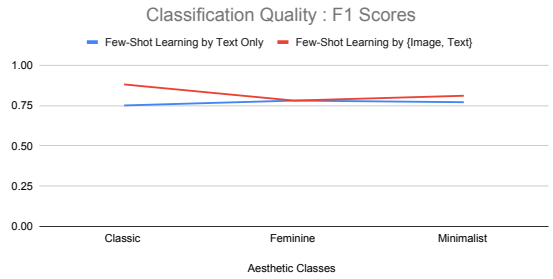

Figure 7: We show the benefit of multi-modality in few-shot learning. Here since we have a smaller number of data samples, we use a line chart that helps show the performance differences between the two lines more clearly.

We can leverage both image and text data in few-shot learning because GPT-4 supports multi-modals. Figure 7 demonstrates the benefits of multi-modality in few-shot learning. Leveraging both image and text data with GPT-4 improves classification performance by 5% to 17%, demonstrating the advantage of GPT-4's multi-modal capabilities.

To conclude, Figure 8 summarizes the classification performance of our proposed approach for all the 18 classes in the testing dataset (Table 1, Figure 4). The weighted F1-score for the entire dataset across all the 18 classes from 3 different categories is 0.802, reaching higher than CaaStle's quality target for the multi-label classification task for every single class and category in CaaStle's dataset. This demonstrates that our new approach is robust.

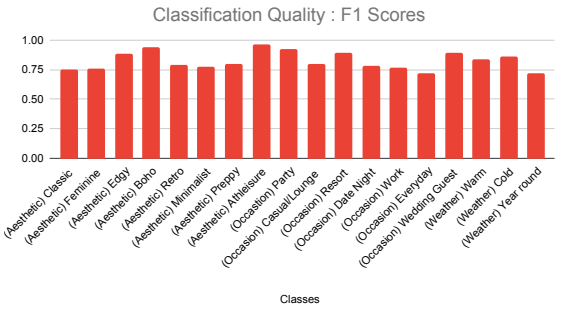

Figure 8: We show the classification performance for all 18 classes in our dataset.

5 CONCLUSIONS

In this paper, we introduced a novel approach that integrates the strengths of LLMs, few-shot learning, and zero-shot learning to create a robust multi-label classification system tailored for the fashion industry. By generating detailed descriptions of new classes and using them as guidelines, our system ensures accurate and scalable classification, adapting seamlessly to the dynamic nature of fashion trends with minimal data requirements. This innovative methodology significantly enhances the efficiency and effectiveness of multi-label classification for fashion items.

Our approach is the first to combine these advanced techniques to address the unique challenges of fashion classification. Through the integration of OpenAI's GPT-4, a state-of-the-art pre-trained LLM, we demonstrated substantial improvements in classification performance, particularly in scenarios with limited labeled data. The few-shot learning process, supported by GPT-4, generates precise class descriptions, which are crucial for effective zero-shot learning. This enables the system to classify new and existing fashion items accurately, maintaining high performance despite the constant influx of new styles and labels.

Additionally, GPT-4's multi-modal capabilities, which allow it to process both image and text data, contribute to the superior performance of our classification system. By leveraging these features, we observed significant improvements in weighted F1-scores across various fashion categories.

This multi-label classification system has already made significant contributions to CaaStle's merchandising and operations. The rapid development of automated, high-quality classification has provided CaaStle with rich semantic data about its inventory, enhancing product capabilities in inventory management, optimization, and personalization. Our approach offers a scalable, flexible, and highly accurate solution, paving the way for further advancements in the fashion industry and beyond.

ACKNOWLEDGEMENTS

We would like to extend our sincere gratitude to the entire Merchandising Team at CaaStle, with special thanks to Francesca De la Rama, Stephanie Shum, Daphne Shapir, Palley Jackson, and Gianni Fuller, for their dedicated efforts in data generation, cleanup, and preparation for the inventory classification project. Their tireless guidance and support have been instrumental in making this work possible.

REFERENCES

Abbas, W., Zhang, Z., Asim, M., Chen, J., and Ahmad, S. (2024). Ai-driven precision clothing classification: Revolutionizing online fashion retailing with hybrid two-objective learning. *Information*, 15(4):196.

Abd Alaziz, H. M., Elmannai, H., Saleh, H., Hadjouni, M., Anter, A. M., Koura, A., and Kayed, M. (2023). Enhancing fashion classification with vision transformer (vit) and developing recommendation fashion systems using dinova2. *Electronics*, 12(20):4263.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Al-Rawi, M. and Beel, J. (2020). Towards an interoperable data protocol aimed at linking the fashion industry with ai companies. *arXiv preprint arXiv:2009.03005*.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020). Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ferreira, B. Q., Costeira, J. P., and Gomes, J. P. (2021). Explainable noisy label flipping for multi-label fashion image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3916–3920.

Guo, S., Huang, W., Zhang, X., Srikhanta, P., Cui, Y., Li, Y., Adam, H., Scott, M. R., and Belongie, S. (2019a). The imaterialist fashion attribute dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

Guo, W., Wang, J., and Wang, S. (2019b). Deep multimodal representation learning: A survey. *Ieee Access*, 7:63373–63394.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Inoue, N., Simo-Serra, E., Yamasaki, T., and Ishikawa, H. (2017). Multi-label fashion image classification with minimal human supervision. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2261–2267.

Kadam, S. and Vaidya, V. (2020). Review and analysis of zero, one and few shot learning approaches. In *Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 1*, pages 100–112. Springer.

Kolisnik, B., Hogan, I., and Zulkernine, F. (2021). Condition-cnn: A hierarchical multi-label fashion image classification model. *Expert Systems with Applications*, 182:115195.

Koonce, B. and Koonce, B. (2021). Resnet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Mallavarapu, T., Cranfill, L., Kim, E. H., Parizi, R. M., Morris, J., and Son, J. (2021). A federated approach for fine-grained classification of fashion apparel. *Machine Learning with Applications*, 6:100118.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Q. Ferreira, B. Costeira, J. R. G., Gui, L.-Y., and Gomes, J. P. (2019). Pose guided attention for multi-label fashion image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Saranya, M. and Geetha, P. (2022). Fashion image classification using deep convolution neural network. In *International Conference on Computer, Communication, and Signal Processing*, pages 116–127. Springer.

Seo, Y. and Shin, K.-s. (2019). Hierarchical convolutional neural networks for fashion image classification. *Expert systems with applications*, 116:328–339.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Xhaferra, E., Cina, E., and Toti, L. (2022). Classification of standard fashion mnist dataset using deep learning based cnn algorithms. In *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 494–498. IEEE.

Zhong, S., Ribul, M., Cho, Y., and Obrist, M. (2023). Textilenet: A material taxonomy-based fashion textile dataset. *arXiv preprint arXiv:2301.06160*.