AI-Enhanced Fashion Discovery: Revolutionizing Online Shopping with Personalized, Multimodal Experiences

Yushu Qiu Weiyue Li Daniel Nurieli Michelle Tan

Background and Motivation

We aim to create an integrated user experience that leverages AI to help users shop for fashion online more efficiently. As of 2024, 57% of internet users have purchased fashion items online, and this number is projected to increase by 13% in 2025. The COVID-19 pandemic, along with the evolving buying habits of Millennials and Gen Z, has driven luxury brands to focus on and invest heavily in enhancing online shopping experiences.

The global fashion market, valued at $1.7 trillion, presents an immense opportunity for innovation. By addressing the challenges in online fashion purchasing, we believe AI has the potential to revolutionize how consumers discover and buy clothing. With advancements in large language models (LLMs) and multimodal technology, we can now create more intuitive and effective user experiences, making it easier for users to find the fashion items they want with ease.

Our inspiration stems from our team member's experience in fashion (collaborated with H&M to improve online fashion shopping experience). Our team's expertise in computer vision, natural language processing (NLP), startups, and industry experience from summer internships will contribute to turning this vision into reality.

Scope and Objectives

Online fashion shopping can be time-consuming as users navigate multiple websites and brands, often turning to social media for inspiration. Our goal is to create an AI-powered platform that aggregates fashion items from various brands, allowing users to quickly and easily find matching items without the hassle of endless browsing.

User Experience Examples

- **Text-Based Search:** Input: "Find me an edgy cocktail dress for a birthday party on a New York rooftop." The platform will suggest curated items based on style and occasion.
- **Image-Based Matching:** Input: An image of a dress. Request: "Find me a pair of shoes and earrings that will give me a classy look with this dress." The platform will suggest complementary items that enhance the overall outfit.

Both experiences can be enhanced with a predefined "user taste" profile that we will create through a 1-2 minute signup process. We envision this recommendation engine improving as the user engages more with the app.

With this approach, we aim to simplify the fashion shopping experience, leveraging AI to make finding and purchasing fashion online faster, more personalized, and more enjoyable.

Technical Implementation

To solve this problem, we aim to train the following models:

- A multimodal model that uses shared embeddings of text and images, such as CLIP, to transform between user queries and items.
- A vision model that will be used as a recommendation engine (self-made model).
- An object detection model to identify items in a given image (Yolo v8).

- A language model (LLM) that will break down user input into actionable instructions, assisting in the shopping process.

Source of Data

- **MD-Fashion-1:** A dataset of about 2 million images from e-commerce sites, fashion shows, social media, and other sources, annotated with category labels and bounding boxes for 80 clothing styles and scenes. This will be used to train our item detection model.
- **DeepFashion2:** A comprehensive dataset with 491K images and 801K labeled clothing items across 13 categories, including attributes like scale, occlusion, viewpoint, and dense landmarks. It contains 873K pairs of commercial-consumer clothing items and is split into training (391K), validation (34K), and test (67K) sets. This dataset will be used for detection and style understanding.
- **Web Scraping:** Images scraped from the internet based on fashion styles and brands, processed with object detection algorithms. These will help build proprietary models for fashion matching, style understanding, and populating a clothing catalog to showcase model functionality.

Research and Development

We hope to build our project on top of:

- **Fashion-CLIP** [1]
- **Yolo-V8** [2]
- **An open-source LLM**
- **Pinecone:** A vector database for storing a vast amount of images for efficient vector search.

Fun Factor

We are all passionate about using multimodal models and tackling a difficult problem with many AI-related components.

Limitations and Risks

Building an accurate model for style recommendation is difficult. Data for the model training may contain noise, and thus requires large amounts of data to work well. Fashion CLIP is not fully expressive and may not fully understand all the text inputs we will use.

Milestones

- Data collection and storage: Build fashion library with data to train our style model and store it in a cloud platform [Oct 4].
- Computer vision model: Train an object detection/segmentation algorithm for extracting items from images [Oct 11].
- LLM training: Fine-tune a shared embedding for images and text such as CLIP; fine-tune an LLM to understand prompts related to fashion and use text to find fashion items [Oct 25].
- Model deployment: Deploy the model and integrate parts together [Nov 8].
- UI design: Create a demo with a UI page that integrates machine learning models [Nov 22].
- Final testing and deployment [Dec 8].

References

[1] P. J. Chia, G. Attanasio, F. Bianchi, *et al.*, “Contrastive language and vision learning of general fashion concepts,” *Scientific Reports*, vol. 12, no. 1, p. 18 958, Nov. 2022, ISSN: 2045-2322. DOI: `10.1038/s41598-022-23052-9`. [Online]. Available: `https://doi.org/10.1038/s41598-022-23052-9`.

[2] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, “A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas,” *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680–1716, Nov. 2023, ISSN: 2504-4990. DOI: `10.3390/make5040083`. [Online]. Available: `http://dx.doi.org/10.3390/make5040083`.