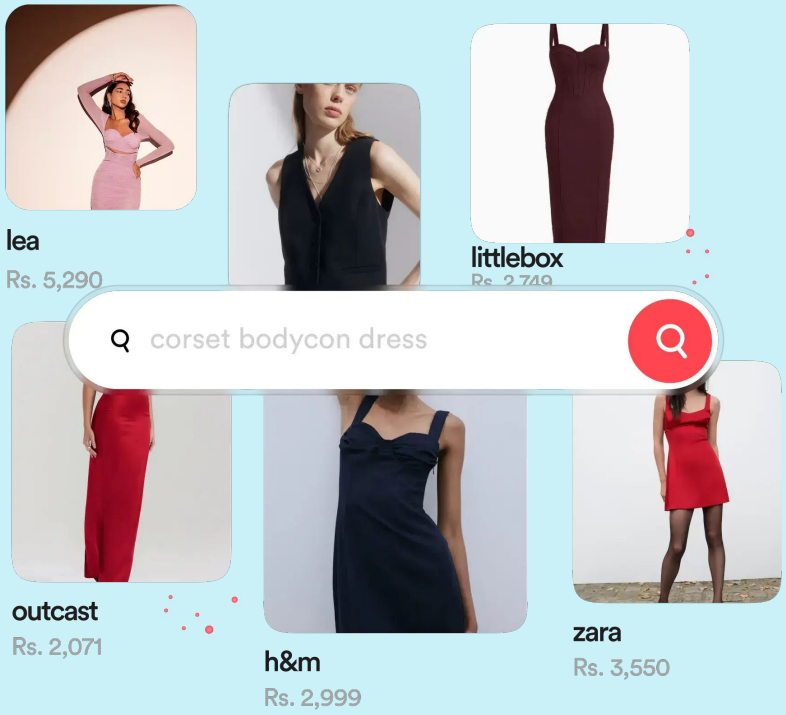
NLP FASHION SEARCH

Kashyap J

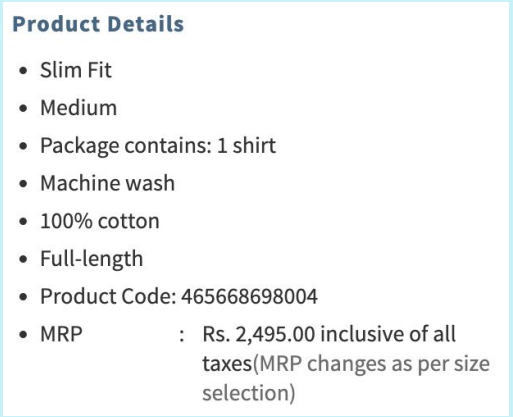


*the real project might not look like this

Recap

GOAL : Automate Fashion description generation for e-commerce and brands.

WHY? : It is necessary for a better search experience and cataloging. Helps with better recommendations.



Product Details

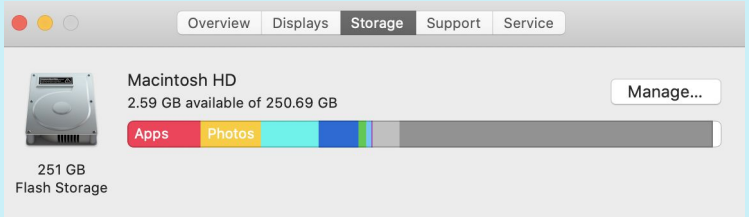
- Slim Fit
- Medium
- Package contains: 1 shirt
- Machine wash
- 100% cotton
- Full-length
- Product Code: 465668698004
- MRP : Rs. 2,495.00 inclusive of all taxes(MRP changes as per size selection)

eg
Poor Product Description

Data

- FashionPedia (open source western clothing dataset)
- Indofashion (open source Indian clothing dataset)

Both the datasets had 30k+ images across categories.
I created a sandbox set of 2k images from each set using random sampling and extracted features/attributes from the metadata + images.



MetaData

sandbox > {} label_descriptions.json > [ ] categories


```
{
  "categories": [
    {
      "id": 0,
      "name": "shirt, blouse",
      "supercategory": "upperbody",
      "level": 2
    },
    {
      "id": 1,
      "name": "top, t-shirt, sweatshirt",
      "supercategory": "upperbody",
      "level": 2
    },
    {
      "id": 2,
      "name": "sweater",
      "supercategory": "upperbody",
      "level": 2
    },
    {
      "id": 3,
      "name": "cardigan",
      "supercategory": "upperbody",
      "level": 2
    },
    {
      "id": 4,
      "name": "jacket",
      "supercategory": "upperbody",
      "level": 2
    },
```


image_id	class_label	brand	product_title
3826	lehenga	Chhabra 555	Peach-Coloured & Golden Sequinned Made to Measure Lehenga & Choli with Dupatta
5230	nehru_jackets	Ethnix by Raymc	Mens Regular Fit BundiRTUA00615-B9100, fancy blue, l
7296	women_kurta	Aks	womens Kurta Sets
2785	kurta_men	DEYANN	Men Navy Blue & White Solid Kurta with Pyjamas & Nehru Jacket
2430	gowns	ADESA	Red A-Line Cotton Striped Kurti for Women
1015	dhoti_pants	Revolution	Plus Size Women's White Solid Loose Fit Trousers
2935	kurta_men	RARE RABBIT	Men White Woven Design Straight Kurta
2840	kurta_men	The Punjabi Gha	Handmade Men's Cotton Kurta
2082	gowns	Ultimate Ecomm	Women's White Color Anarkali Embroidery Work Salwar Suit
2595	kurta_men	GRACIT	Men Black & White Solid Kurta with Pyjamas
1170	dhoti_pants	N/A	dhrona Solid Men Dhoti
7377	women_kurta	OM KALYANAM	Women Cotton Key Print Kurti With Palazzo Set (KP58_Grey_NavyBlue_S)
694	blouse	Master Weaver	Boat Neck Ikkat and South Cotton Short Sleeves Blouse with Hand Embroidery Mirrors

Exploration

Image: 0f6ff6066489017c55664bcc6d56ff6f
Annotations: 4

Image: 8fb5c16e87c1d2f8bcfa784c0cd1a4bf
Annotations: 13

Image: 90da28d23c2600b7845fbbcc050b852a
Annotations: 6

Attribute Prediction

Fine-tuned Resnet-50 to learn the attributes that the images have. I used the images, the metadata provided and multi-hot attribute vectors (encoding). Then used the model to predict attributes.

Evaluation
Macro Precision: 0.2605
Macro Recall: 0.1511
Macro F1: 0.1731
Micro F1: 0.6484

Inference
Performance on rare attributes is poor.
A lot of false positives especially for attributes that appear >5 times in the dataset.

Fix
Balancing the dataset
And maybe regularisation to improve learning of rare attributes

Description Generation

Using the attributes, the images and the metadata I created prompts to generate rich text descriptions.

I tried 2 combinations

- BLIP-2 (Vision LM, feature extraction) + Flan T5 (LLM)
- LLaVA (Vision LM,feature extraction) 13b Model

Description Generation

I used the same prompts for both combinations.

"USER: <image>\n"
"Describe this clothing item in detail. Do not mention words like lady, woman, man, model, etc. "
f"It has the following characteristics: {attr_str}. Focus on its visual appearance, including the color, material, any patterns, the style of the garment (e.g. casual, formal) and specific features like the neckline, sleeves, etc and overall silhouette. "
"ASSISTANT:"

Description Generation

BLIP-2 + Flan T5 << LLaVA

Hypothesis as to why LLaVA outperformed the other combination

1. Better feature attribute extraction from the images (metric - BLIP-2 printed attributes that didn't exist in the image or were of minor importance, LLaVA chose the top 3 features)
2. LLaVA's generations were more structured and complete. Flan T5 mentioned "features a lady" , "the model", etc despite being asked not to.

I chose to go with LLaVA

Maybe changing the prompt would help and having separate prompts for the two datasets could improve LLaVA's outputs.

“Prompt Engineering”

For FashionPedia

(same prompt)

"USER: <image>\n"
"Describe this clothing item in detail. Do not mention words like lady, woman, man, model, etc. "
f"It has the following characteristics: {attr_str}. Focus on its **visual appearance**, including the color, material, any patterns, the style of the garment (**e.g. casual, formal**) and specific features like the neckline, sleeves, etc and overall silhouette. "
"ASSISTANT:"

For IndoFashion

(same prompt)

"USER: <image>\n"
"You’re an e-commerce copywriter. Describe this garment in vivid detail."
"focus on its color, fabric/material, pattern or **embroidery,** silhouette, "
"**neckline, sleeves or drape style**, length, and any **special accents or embellishments**."
f"This is {descriptor}. "
"Do **NOT** mention model, person, or background, only describe the garment itself.\n"
"ASSISTANT:"

Challenges

Incomplete descriptions generated

- The neckline is scoop-style, and the tank top is hanging on a wooden hanger. The material appears to be lightweight and comfortable, making it

This was because of the max-token size constraint of the model was met.

Solution : Fixing it in post - ensuring that <EOS> token was reached so that the output was well-formed

Prompt Sensitivity

Small variations in the prompt : "describe this garment" vs "describe this piece of clothing" changed the style and focus of output. Sometimes, it also lead to features not being present also being mentioned.

This was because of the max-token size constraint of the model was met.

Fix: Iterated and tested multiple prompt templates before choosing the final ones for Fashionpedia and IndoFashion.

Challenges

Intent / Use Case mentions.

- Descriptors like festive, for a party or office-ready were rarely generated because the model lacked event based context.

Including that in the prompt helped increase event/intent mentions.

Another possible solution would be to explore conditioning on user intent or textual metadata in multi-modal setups.

Unclean Metadata

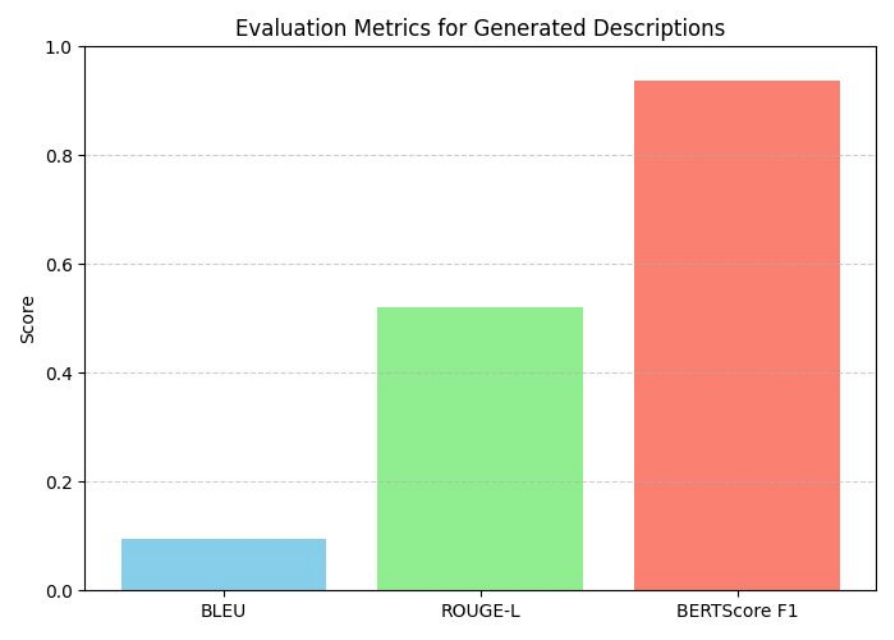
Indofashion particularly had a lot of brand names, non-english words and inconsistent labels. This was because all the data was scraped from sites like Myntra, Meesho, etc.

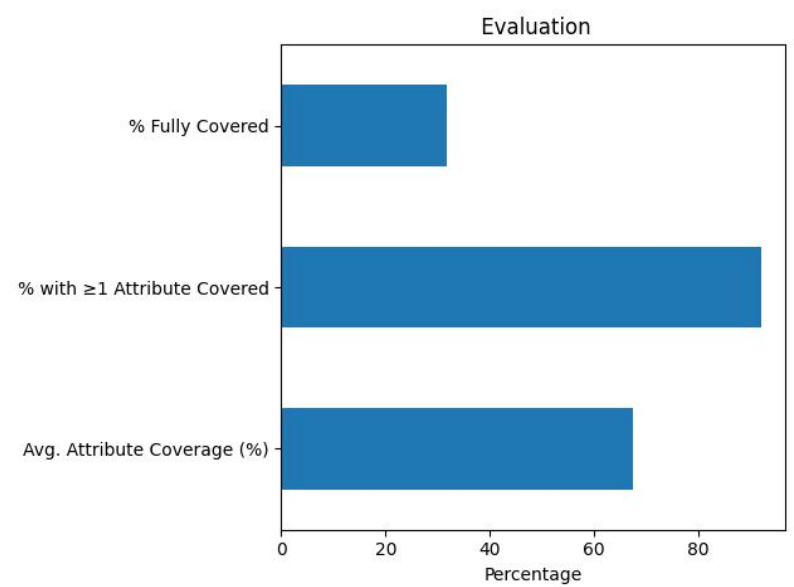
"DHRUVI TRENDZ Women's Anarkali Kurta" , "Honey Fashion Solid Sherwani"

This required cleaning and ensuring that only the important features remained + influenced the output.

Evaluation

Asked friends to describe in detail some of the images (~100) and compared those with the LLaVA generated descriptions.





Embedding + Retrieval

- After the descriptions have been generated, I used a sentence transformer model - MiniLM to convert text descriptions into numerical vector representations (embeddings). Also Normalised the embeddings to have unit length.
- Created a FAISS index (Facebook AI Similarity Search) using the embeddings. The user query is also converted into the same index and then a similarity matching using dot product of the vectors.

Retrieval Performance:

P@1: 0.998, R@1: 0.998

P@5: 0.200, R@5: 1.000

P@10: 0.100, R@10: 1.000

Basically, 1/5 images matches the query (most of the times)

DEMO?

Key Observations and Learnings

- Prompting technique makes a huge difference.
- Metadata cleaning was critical, especially for noisy datasets like IndoFashion.
    - Metadata can be generated using Vision Models like CogVLM but it is time consuming and requires large amount of compute
- It could have been a better idea to focus on one dataset and improve the model's performance through hyper-specific prompting and metadata
- Not all fashion attributes are equally easy to classify, some attributes (like Pattern) performed well while others seemed to require more fine-tuning (maybe use LoRA).

Future Improvements/Possibilities

- Image search - Input image and find something similar
  - There's startups that are doing this now - https://shoppin.app/
- Include more images in the dataset - the more the merrier.
- Create specialised embedding subspaces for different attribute categories
- Better understanding of user intent. It can do better with event/season based queries.

Thank you!