Fashion Product Identification and Categorization

Rationale behind choice of use case.

In today's world of online shopping and social media influence, the fashion industry and its technologies are constantly evolving. There is an increasing number of people who turn to digital platforms for their shopping needs, whether online websites (Zalando) or social media shops (Instagram or TikTok). With such an increase in demand, the ability to quickly and accurately identify and analyze fashion items and trends has become essential. Therefore, the case that I decided to focus on for this assignment was to use a transformers model for 'Fashion Product Identification and Categorization' of fashion images derived from online sources.

In class, we have experimented with different models ranging from NLP to visual tasks, but in this task, I wanted to focus on image classification as it is a new area of knowledge for me and I would like to improve my skills in it. Apart from this, seeing the increase in demand for online shopping in our current generation, I wanted to explore a use case that is relevant to today's industries. I will be specifically using ViT (vision transformers) to tag and categorize fashion items automatically. Such automation could lead to an improvement in the online shopping experience through better recommendations for users, leading to happier customers and more sales. Additionally, it could help businesses manage their inventory more efficiently by automatically updating products in stock and reducing wasted resources.

Moreover, fashion product identification can provide valuable insights into current fashion trends. By analyzing patterns in the data, businesses can stay ahead of the curve and adjust their strategies accordingly. In short, this project can be a starting point for experimenting with the way we shop online and how the fashion industry operates.

Potential advantages and applications of your chosen use case.

Using a transformer model for fashion analysis offers several advantages. As previously mentioned, it can improve the online shopping experience by providing personalized recommendations, leading to happier customers and an increase in sales. Secondly, it can help businesses manage their inventory more effectively, reducing costs, minimizing waste, and being more sustainable - which is highly valued in the current market. Thirdly, it gives companies valuable insights into what's trending in fashion, helping them stay competitive.

Currently, there exist AI start-ups such as 'Styl', a recently released platform that offers personalized recommendations based on what items you 'swipe' left or right - similar to Tinder for clothes. This start-up started to gain attention online by publishing videos on TikTok and offering an easier way to online shop, without having to endlessly scroll through fashion websites. There have been other incentives such as identifying fashion trends based

on object identification from runway shows, depicting the potential applications of this use case.

Detail the datasets you are using, including why they were chosen.
Initially, I started working with the 'DeepFashion' dataset found on HuggingFace which consisted of 100k rows, three different columns of images, and various columns of detailed information, such as mask, mask_overlay, or position of the model. This dataset also did not have labels for images but rather descriptions, which caused complications on my side. It required a lot of preprocessing before it was ready to be fed into the model, and even after the preprocessing, there were issues with the training of the model which would not run. After many trials, I decided to look for other more well-known datasets such as 'Fashion-MNIST', which is commonly used as a benchmark for machine learning algorithms.

'Fashion-MNIST' consists of two columns, the image and the label, which include 10 different categories of clothing. Different from the first dataset, the latter was cleaner and simpler to preprocess with fewer unnecessary columns and details. While I still had to experiment a lot to preprocess images in the correct format for my vision transformer model, this dataset allowed for faster iterations. Hence, 'Fashion-MNIST' was the better choice because it was better organized and it allowed me to focus more on the model training and tuning without the overhead of extensive data cleaning.

Explain why you selected the transformer model(s) for your project.
I chose a Vision Transformers (ViT) model because it is suited for understanding the global context of images, and hence widely used for tasks such as image classification. Transformer models are also better at capturing these details better than others, such as CNNs, which is why they were selected for this project. Though this assignment is transformer-based, I tried out both ViT and CNNs to compare their performance, something I mentioned in the alternative models I experimented with. Further, we have worked with different examples of ViT in class from which I got an overall understanding of how it should work. As a newly learned model, I wanted to explore a use case in which I could use ViT but simply implement it in another domain.

Describe any alternative models you considered or experimented with.
Other models I considered were BERT, GPT, and BART, but in other use cases. As I was brainstorming ideas, I considered working on legal document summarization, basic image classification, and style transfer. As I was looking at what models I could implement for each case, I decided to go for image classification but in the fashion domain.

Before doing a ViT model for the task, I initially tested a convolutional neural network (CNN) as I was more familiar with its structure and wanted to look at its performance. The results were satisfactory with an accuracy of 91%, but given our project's requirements, I

moved on to the transformers model. I researched online for different transformer variants and looked into ViT, DeiT, and Swin Transformers. I decided to focus on ViT because of my familiarity with it and its approach to processing images in chunks, similar to words in a sentence, and it seemed ideal for this task.

Instead of trying too many different models, I experimented a lot with the fine-tuning of the ViT model and transformations of the dataset so the model would successfully train on my data. The biggest challenge was transforming images to a format readable by the ViT model, as it is a task not very familiar to me yet. Multiple trials resulted in either an error in training or a crash due to a lack of GPU resources. After multiple trials and errors, I achieved a model successful at identifying fashion products, with an accuracy of 94.35%.

Explain how you are assessing your model's performance.

To evaluate the model, I used several performance metrics: accuracy, precision, recall, and F1 score macro-averaged. These metrics help us understand how well the model is categorizing the fashion items in the images of the dataset. Accuracy, the overall correctness of the model, was found to be 94.35%, and precision, recall, and the F1 spcre were 0.944, 0.943, and 0.943, respectively. The other evaluation used was a confusion matrix to determine the performance of the model in classifying different categories of clothing. From the results it was seen that one area for improvement would be the model's classification of shirts and T-shirts, which ViT had difficulties differentiating between potentially due to their similarities. Apart from that, the overall results show that the model is very accurate in identifying and categorizing fashion items, therefore it can be the starting point a real-life use case.