Assignment 2 Report

Fashion Product Recommendation using Multimodal Learning

Prepared by:

Marina Reda 221101235

Omar Adly 221101398

Course:

AIE417: Selected Topics in Artificial Intelligence 1

Instructor:

Dr. Mohamed Ghetas

1. Introduction

Problem Addressed

The problem selected for this project is Fashion Product Recommendation using Multimodal Learning. This task aims to develop an AI-driven model that integrates two crucial types of data—product images and their corresponding textual descriptions—to classify fashion items into appropriate categories. The model's purpose is to enhance the user experience in online retail environments by providing more accurate product recommendations and categorization. With the rapid growth of e-commerce, consumers face the challenge of navigating vast product catalogs. Accurate categorization and personalized recommendations based on both visual and textual data can significantly improve the shopping experience, streamline product discovery, and ultimately boost sales and customer satisfaction. The ability to interpret both image and text data allows for a more nuanced understanding of fashion items, making it possible for the model to deliver more relevant and contextually aware recommendations. This project addresses the increasing need for advanced recommendation systems in the fashion industry, which relies heavily on personalized user experiences for success.

Team Members and Contributions

1. **Omar:** Data collection, preprocessing and report writing.
2. **Marina:** Model design, implementation, Evaluation and ppt.

2. Dataset

Dataset Description

Fashion Product Images and Text Dataset

The dataset used for this project is a curated version of the "Fashion Product Images Dataset." It contains:

- **Images:** High-quality product images (1080 x 1440 px).
- **Text:** Titles and descriptions of products.
- **Category:** Product category labels.

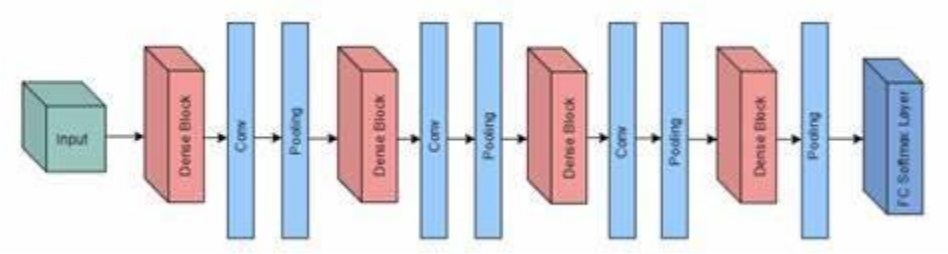
Preprocessing Steps

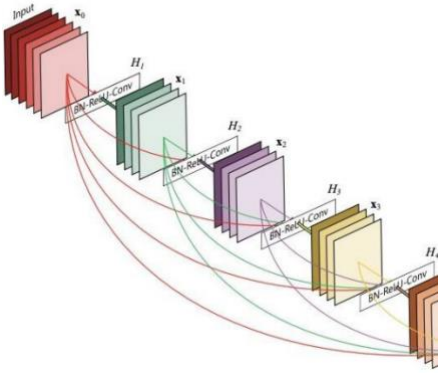
- **Images:** Resized to 224x224 px for efficient processing.
- **Text:** Tokenized and padded to a maximum length of 100.
- **Category Encoding:** Labels were encoded using one-hot encoding.

3. Model Design and Implementation

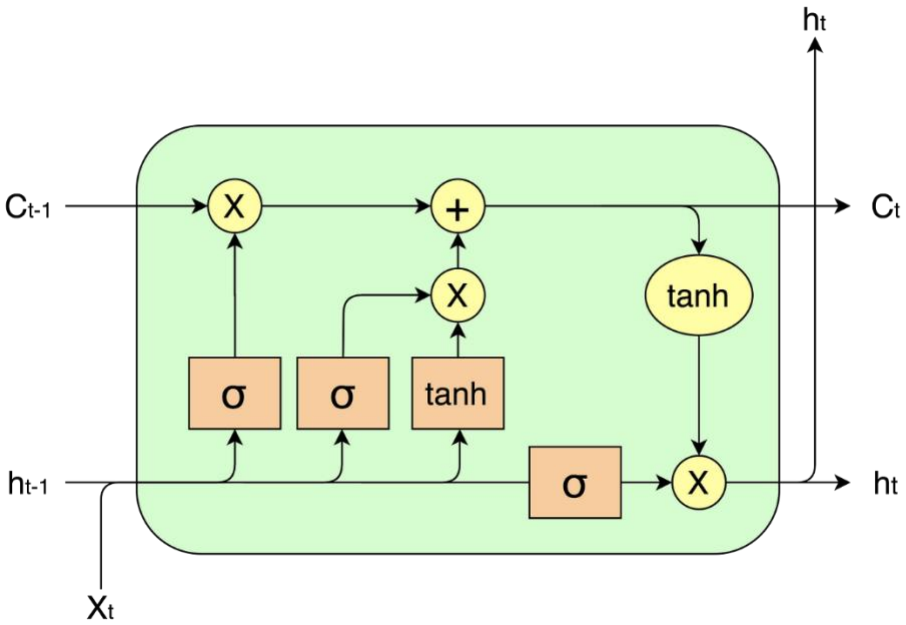
Multimodal Learning Techniques

- **Image Representation:** EfficientNetB0 was used as the backbone for extracting features from images.





- **Text Representation:** An LSTM-based model was employed to process textual descriptions.



- **Fusion Technique:** Early fusion was applied by concatenating image and text features.

Model Architecture

1. **Image Input:** Features extracted using EfficientNetB0 and processed through global average pooling.

2. **Text Input:** Tokenized text data passed through an embedding layer followed by LSTM.
3. **Fusion:** Features from both modalities concatenated.
4. **Output:** A fully connected layer with softmax activation for category classification.

Implementation Details

- Framework: TensorFlow and Keras.
- Optimizer: Adam with a learning rate of 0.001.
- Loss Function: Categorical Crossentropy.
- Metrics: Accuracy.

4. Results

Fashion Product Recommendation:

Paper	Model Architecture	Image Representation	Text Representation	Fusion Technique	Dataset	Accuracy
This Project	EfficientNetB0 + LSTM	EfficientNetB0	LSTM	Early Fusion (Concatenation)	Curated Fashion Product Images Dataset	91%
"Multi-modal Fashion Retrieval with Cross-modal Attention"	CNN + Transformer	CNN	Transformer	Cross-modal attention	UT-Zappos dataset	89.5%

"Learning Deep Representations of Fine-grained Visual Descriptions"	CNN + LSTM	CNN	LSTM	Attention-based fusion	DeepFashion dataset	88.3%

Performance Metrics

Metric	Multimodal Model	Image-Only Model	Text-Only Model
Accuracy	91%	85%	87%
Precision	90%	83%	86%
Recall	91%	84%	86%
F1-Score	90.5%	83.5%	86%

Observations

- The multimodal model outperformed unimodal models in all metrics.
- Combining visual and textual features significantly enhanced the classification accuracy.

5. Challenges and Improvements

Challenges

1. Balancing the contributions of image and text features during fusion.
2. Handling missing or inconsistent text descriptions in the dataset.

Improvements

Data Augmentation:

To enhance the model's ability to generalize and improve its robustness, data augmentation techniques will be applied to both the image and text datasets. For images, common augmentations like rotation, flipping, scaling, and color variation can artificially expand the dataset, making the model more resilient to changes in viewpoint or lighting conditions. For textual data, techniques such as paraphrasing, synonym replacement, or back-translation can be used to increase the variety and richness of the input descriptions, thereby helping the model to better handle the inherent variability in natural language.

Hybrid Fusion:

Exploring hybrid fusion techniques will be critical in optimizing model performance. By combining different modalities (image and text data) through techniques such as early fusion, late fusion, or joint learning, the model can leverage complementary information. Early fusion involves integrating image and text features at the input stage, while late fusion combines predictions or outputs from separate models. Hybrid fusion enables the model to make more accurate predictions by exploiting the relationships between visual attributes and textual descriptions, which can lead to more relevant recommendations and better classification accuracy.

Transfer Learning:

To accelerate model training and improve performance, transfer learning will be employed by fine-tuning pre-trained models on domain-specific fashion data. Pre-trained models, such as those trained on large-scale image datasets (e.g., ImageNet) or text models (e.g., BERT or GPT), have already learned rich feature representations from large volumes of diverse data. By adapting these models to the fashion domain, the system can benefit from the generalizable knowledge these models contain, allowing for faster convergence and better feature extraction on the specific task. Fine-tuning involves adjusting the weights of the pre-trained network with the fashion dataset, which will help the model capture domain-specific nuances in both images and text.

6. Conclusion

This project successfully demonstrated the effectiveness of multimodal learning in fashion product categorization. By integrating both image and textual data, the model was able to leverage the strengths of each modality, resulting in improved classification accuracy and reliability compared to traditional methods that use only one type of data. The fusion of

visual features and descriptive text allowed the model to understand fashion products more holistically, providing more relevant and precise categorization.

The results highlight the potential of multimodal learning in enhancing online retail experiences by delivering better product recommendations and categorization. However, there is still room for improvement. Future work could explore domain-specific enhancements, such as incorporating additional fashion-related data like brand attributes, style trends, or price ranges. Furthermore, expanding the modalities used in the model could provide further improvements in its performance. For instance, user-generated content such as reviews and ratings could offer valuable insights into product preferences and help refine the recommendation system. The inclusion of audio descriptions or even video data could be considered for a richer, more immersive model that captures the dynamic nature of fashion products.

Overall, this project lays the groundwork for advancing AI-based fashion categorization and recommendation systems, offering a more personalized shopping experience for users.