Deep Learning

E-Commerce Recommendations via Multi-Modal Inputs

Team 3
Rachel Brooks & Eric Brown & Claudio Mema

School of Graduate Studies

Data Science
DAAN 570 – Deep Learning
Fall Semester, 2024

Document Control

Work carried out by:

Name	Email Address	Exhaustive list of Tasks
Rachel Brooks	reb5653@psu.edu	
Eric Brown	esb191@psu.edu	
Claudio Mema	cqm6179@psu.edu	

Revision Sheet

Date	Revision Description

TABLE OF CONTENTS

INTRODUCTION

In the rapidly evolving landscape of e-commerce, personalized shopping experiences are crucial for attracting and retaining customers. Imagine stepping onto a vibrant online website where your wish list shows products tailored just for you. As you scroll through the site it seems to know your tastes, preferences, and even your mood better than your close known friend. Guiding you to treasures you never knew you wanted. This is the magic of an adaptive e-commerce recommendation system—a digital shopping companion that transforms the online shopping experience from overwhelming to delightful. A system that not only improves customer satisfaction but also drives sales and increases conversion rates. A system that serves as a vital tool that enhances user engagement by suggesting relevant products based on individual preferences and behaviors. While deep learning technologies have been incorporated in e-commerce platforms such as Amazon, the exploration of real-time analysis on mood, seasonal trends, and social influence is an area still being explored to make a more dynamic and intuitive experience for customers. That is where our e-commerce recommendation system hopes to shine. Instead of relying solely on past purchases and browsing history, our recommendation system would use a multi-modal data integration approach incorporating various data types. Using product images, user reviews, social media activity, and even location data, we hope to create a richer, more nuanced understanding of user preferences. This approach allows for recommendations that feel more like a conversation than a transaction.

PROBLEM STATEMENT

The primary objective of our project is to develop a multi-modal recommendation system that utilizes deep learning techniques to enhance the accuracy and relevance of product recommendations in an e-commerce platform. By combining diverse data types—such as images, textual descriptions, and contextual information—this system aims to create a comprehensive understanding of user preferences and behaviors.

CHALLENGES

Developing a multi-modal data integration system for an e-commerce recommendation platform presents several challenges, especially when creating a proof of concept (PoC). One significant hurdle is the cold start problem, which complicates the provision of accurate recommendations for new users or products with limited interaction data. Scalability is another critical concern, as the system must efficiently handle vast amounts of data and process real-time updates. The integration of diverse data sources while ensuring high data quality is complex, particularly due to the high dimensionality of the data, where certain features may not provide valuable insights for categorizing user preferences. This complexity can hinder explainability, making it difficult to justify recommendations.

Additionally, designing a robust deep learning architecture that effectively fuses information from images, text, and user interactions requires substantial computational resources and meticulous model design. While, ongoing challenges like sparse data, managing the exploration-

exploitation trade-off, and mitigating biases in recommendations further complicate the process. User trust and privacy are also paramount, as striking a balance between maintaining user privacy and collecting sufficient data for personalization is delicate. Finally, accurately evaluating the system's performance—considering both immediate user satisfaction and long-term business metrics—poses its own set of challenges in this dynamic environment.

RELATED WORKS

A significant body of research has explored the application of artificial intelligence and deep learning techniques in e-commerce recommendation systems, which are essential for providing personalized product suggestions across various categories such as electronics, fashion, and home goods. Zhang et al. conducted a comprehensive survey analyzing over 100 research papers on deep learning-based recommender systems, highlighting various neural network architectures utilized in this domain. Our review of key studies indicates that most models are rooted in neural network architectures, particularly leveraging collaborative filtering and content-based approaches, while many researchers have embraced hybrid models to enhance recommendation accuracy.

Notably, Cheng et al. proposed the Wide & Deep learning model, which integrates a wide linear model with a deep neural network to capture both memorization and generalization in recommendations—an approach that industry leaders like Google currently adopt for app recommendations. Similarly, advanced multi-modal systems have emerged, integrating diverse data types such as product images, descriptions, and customer reviews. For instance, Wang et al. (2019) presented a multi-modal recommendation system that enhances suggestions in the electronics category, while Zhang et al. (2020) introduced a furniture recommendation system that combines visual and textual features to align user preferences with product aesthetics and functionality.

Several studies have focused on incorporating sequential information into recommendations. Hidasi et al. utilized recurrent neural networks (RNNs) for session-based recommendations, achieving significant improvements over conventional methods. Building on this work, Li et al. introduced the Neural Attentive Session-based Recommendation (NASR) model, which employs attention mechanisms to capture users' sequential behaviors and intents during sessions. Additionally, Wang et al. proposed the Neural Graph Collaborative Filtering (NGCF) model, leveraging user-item graph structures, while Ying et al. developed PinSage, a GNN-based recommendation system deployed at Pinterest that efficiently generates embeddings for billions of items.

To tackle the cold-start problem, Volkovs et al. combined content and collaborative information to enhance performance for new items and users, while Barkan et al. introduced Item2Vec to learn item embeddings from purchase sequences, proving effective in cold-start scenarios. Efforts to improve interpretability have led to frameworks like Chen et al.'s Attentive Collaborative Filtering, which provides explanations for recommendations, and Ma et al.'s combination of collaborative filtering with knowledge graphs for explainable recommendations.

In addition to these advancements, our project aims to build upon existing works by integrating visual, textual, and potential video inputs into a unified recommendation system tailored for general e-commerce platforms. While prior research has successfully applied these techniques in specific domains, such as fashion and electronics, our system is designed to offer personalized recommendations across all e-commerce categories, making it a flexible and scalable solution. By leveraging the strengths of multi-modal data integration and advanced deep learning techniques, we seek to enhance the relevance and accuracy of recommendations, ultimately improving user engagement and satisfaction. Collectively, these studies illustrate significant advancements in AI-driven e-commerce recommendation systems, emphasizing the promise of neural network architectures and hybrid approaches in enhancing accuracy, addressing cold-start challenges, and improving personalization and interpretability.

IMPORTANCE AND IMPACTS

The research problem addressed in this project focuses on improving the personalization and accuracy of product recommendations in the fashion domain through the use of advanced multi-modal learning models. This research problem is significant because recommendation systems are central to modern e-commerce and customer interaction strategies. By integrating cutting-edge models like CLIP and ResNet, this project showcases how technology can bridge the gap between human preferences and machine understanding, enabling scalable solutions for both businesses and consumers.

DATA COLLECTION

We will use the Fashion Product Images (Small) dataset from Kaggle for this project. This dataset contains 44,443 images of fashion products from six main categories (apparel, accessories, footwear, etc.), along with metadata such as product descriptions, brand information, and price. https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small

- **Format:** The dataset includes images (JPG) and a CSV file with associated metadata.
- **Size:** The total dataset size is approximately 44,443 images, each with dimensions of 80x60 pixels.
- **Preprocessing:** We will normalize the images and resize them to a standard size to be compatible with the input requirements of our CNN model. We will also tokenize and preprocess text descriptions to extract semantic information, which can be used alongside visual embeddings for more personalized recommendations.

Overall Dataframe Overview:

	id	gender	masterCategory	subCategory	articleType	baseColour	season	year
0	15970	Men	Apparel	Topwear	Shirts	Navy Blue	Fall	2011.0
1	39386	Men	Apparel	Bottomwear	Jeans	Blue	Summer	2012.0
2	59263	Women	Accessories	Watches	Watches	Silver	Winter	2016.0
3	21379	Men	Apparel	Bottomwear	Track Pants	Black	Fall	2011.0
4	53759	Men	Apparel	Topwear	Tshirts	Grey	Summer	2012.0
5	1855	Men	Apparel	Topwear	Tshirts	Grey	Summer	2011.0
6	30805	Men	Apparel	Topwear	Shirts	Green	Summer	2012.0
7	26960	Women	Apparel	Topwear	Shirts	Purple	Summer	2012.0
8	29114	Men	Accessories	Socks	Socks	Navy Blue	Summer	2012.0
9	30039	Men	Accessories	Watches	Watches	Black	Winter	2016.0
10	9204	Men	Footwear	Shoes	Casual Shoes	Black	Summer	2011.0
11	48123	Women	Accessories	Belts	Belts	Black	Summer	2012.0
12	18653	Men	Footwear	Flip Flops	Flip Flops	Black	Fall	2011.0
13	47957	Women	Accessories	Bags	Handbags	Blue	Summer	2012.0
14	46885	Boys	Footwear	Flip Flops	Flip Flops	Navy Blue	Fall	2012.0
15	12369	Men	Apparel	Topwear	Shirts	Purple	Fall	2011.0
16	29928	Men	Accessories	Watches	Watches	Black	Winter	2016.0

usage	productDisplayName
Casual	Turtle Check Men Navy Blue Shirt
Casual	Peter England Men Party Blue Jeans
Casual	Titan Women Silver Watch
Casual	Manchester United Men Solid Black Track Pants
Casual	Puma Men Grey T-shirt
Casual	Inkfruit Mens Chain Reaction T-shirt
Ethnic	Fabindia Men Striped Green Shirt
Casual	Jealous 21 Women Purple Shirt
Casual	Puma Men Pack of 3 Socks
Casual	Skagen Men Black Watch
Casual	Puma Men Future Cat Remix SF Black Casual Shoes
Casual	Fossil Women Black Huarache Weave Belt
Casual	Fila Men Cush Flex Black Slippers
Casual	Murcia Women Blue Handbag
Casual	Ben 10 Boys Navy Blue Slippers
Formal	Reid & Taylor Men Check Purple Shirts
Casual	Police Men Black Dial Watch PL12889JVSB

Feature Description:

Feature Descriptions:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44446 entries, 0 to 44445
Data columns (total 12 columns):

#	Column	Non-Null Count	Dtype
0	id	44446 non-null	int64
1	gender	44446 non-null	object
2	masterCategory	44446 non-null	object
3	subCategory	44446 non-null	object
4	articleType	44446 non-null	object
5	baseColour	44431 non-null	object
6	season	44425 non-null	object
7	year	44445 non-null	float64
8	usage	44129 non-null	object
9	productDisplayName	44439 non-null	object
10	Unnamed: 10	22 non-null	object
11	Unnamed: 11	2 non-null	object

dtypes: float64(1), int64(1), object(10)

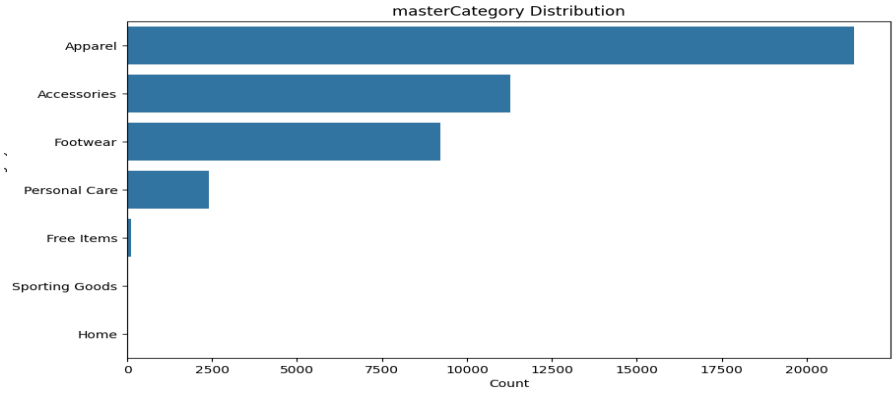
Main columns of interest for model:

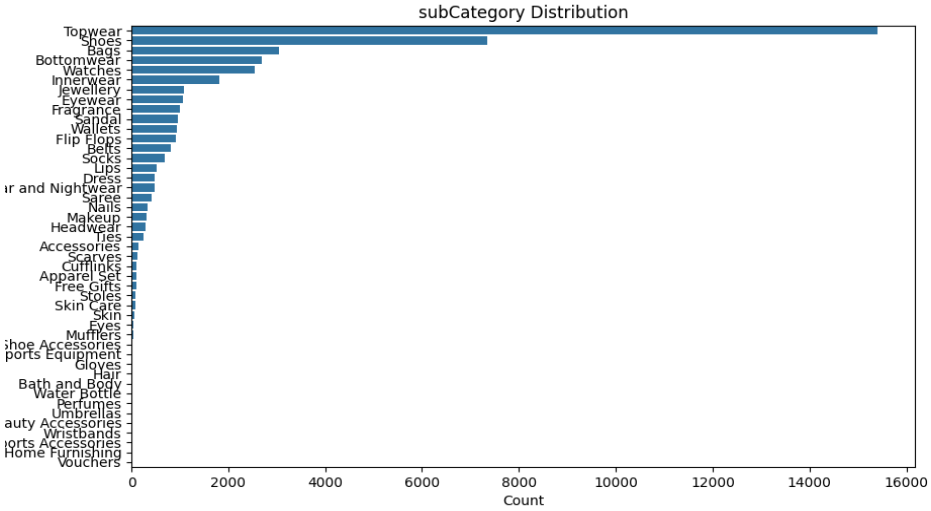
'baseColour', 'masterCategory', 'subCategory', 'season', 'gender', 'articleType'

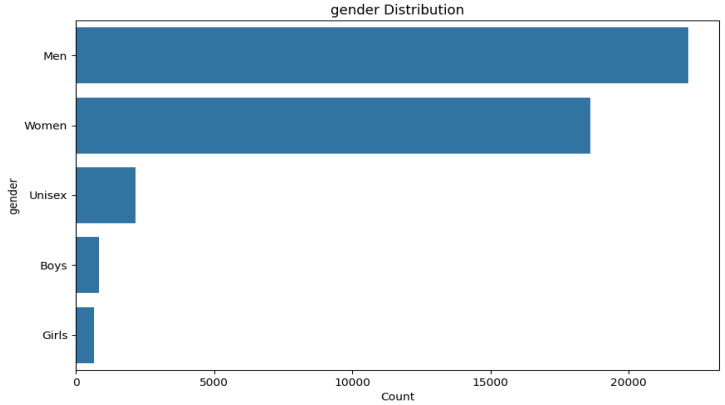
Missing Values:

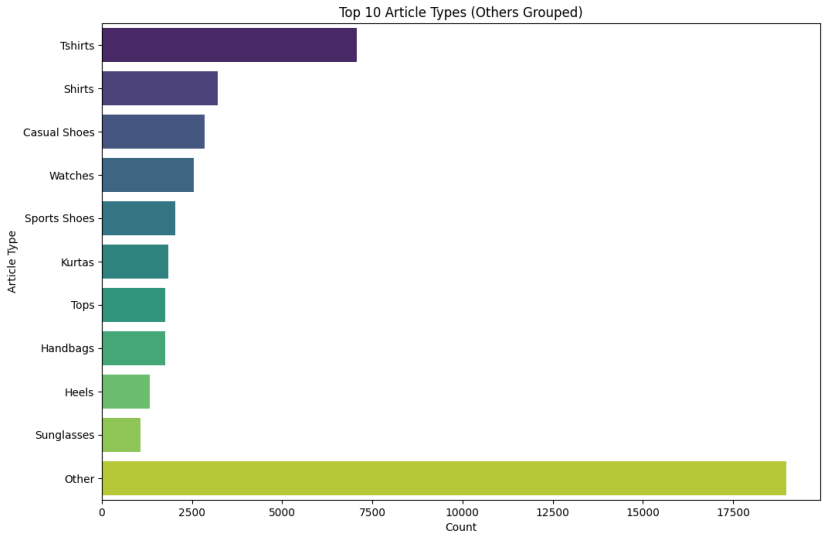
baseColour	15
season	21

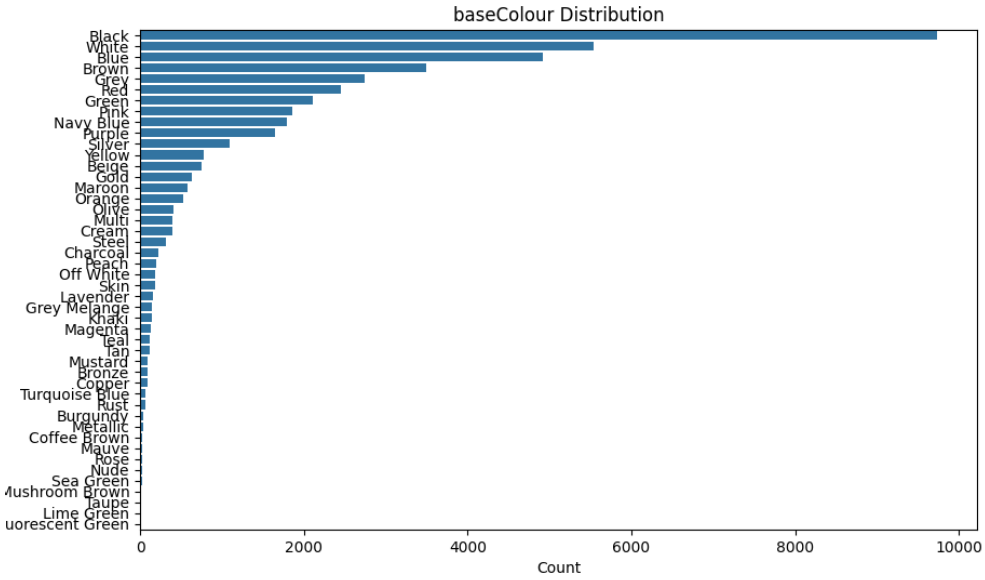
Visualization of Data:











Inferences of Data:

1.) Missing Values

- **baseColour**: 15 missing values indicate that a small proportion of items lack a primary color specification.
- **season**: 21 missing values suggest that some products are not associated with any specific season.
- **Conclusion**: Missing values are minimal but need handling. Options include removing records that have missing baseColour or season data or replacing them with an "Unknown" category.

2.) Feature Distribution

a. Base Colour

- **Dominance**: The most frequent colors are `Black`, `White`, and `Blue`, which are commonly popular in fashion.
- **Diversity**: A wide variety of colors are present, but the tail suggests many infrequent or niche colors.
- **Insight**: We might group rare colors into an "Other" category if they don't significantly contribute to the analysis.

b. Master Category

- **Dominance**: `Apparel` is the most dominant category, followed by `Accessories` and `Footwear`.
- **Minor Categories**: Categories like `Free Items`, `Sporting Goods`, and `Home` have significantly fewer products.
- **Insight**: The dataset is heavily skewed towards apparel, which aligns with our focus on fashion. However, we should probably remove the home category since very few records are listed inside of this field.

c. Subcategory

- **Top Subcategories**: Categories like `Topwear`, `Shoes`, and `Bottomwear` dominate the distribution, reflecting consumer demand for these staples.
- **Long Tail**: A long tail with niche subcategories like `Skin Care` and `Umbrellas`.
- **Insight**: The model might benefit from aggregating less frequent subcategories into broader groups.

d. Gender

- **Dominance**: `Men` and `Women` categories are well represented, with `Unisex` making up a smaller proportion.
- **Minor Groups**: `Boys` and `Girls` categories are underrepresented, potentially limiting generalizability for children's fashion.
- **Insight**: Gender-based segmentation is feasible, but limited representation of children's categories might impact predictions for those demographics.We will probably have to assign weights to this field to better represent it.

e. Article Type

- **Top Article Types**: Items like `Tshirts`, `Shirts`, and `Casual Shoes` dominate the distribution.
- **Other Items**: A significant number of less frequent article types grouped as "Other."
- **Insight**: The high dominance of a few types suggests focusing on these for training. Rare types might be grouped to avoid data sparsity issues.

f. Season

- **Seasonal Spread**: The distribution is relatively balanced across `Summer`, `Winter`, and `Fall`.
- **Insight**: Seasonal trends can be analyzed further for recommendations (e.g., `Summer` products for warmer climates).

3.) Relationships and Patterns

Gender vs Article Type

- **Observation**: Certain article types are gender-specific (e.g., Sarees for women and Blazers for men), while others like Tshirts or Shoes are more evenly distributed.
- **Insight**: Gender strongly influences product preferences, suggesting its importance as a feature for personalization.

Season-wise Distribution by Gender

- **Observation**: Men and women seem to have fairly balanced seasonal preferences, with slight variations.
- **Insight**: This can guide the model to account for seasonality in recommendations.

DATA PREPROCESSING

Total Missing Values Including Attributes of Non-Interest

year	1
usage	317
productDisplayName	7
Unnamed: 10	44424
Unnamed: 11	44444

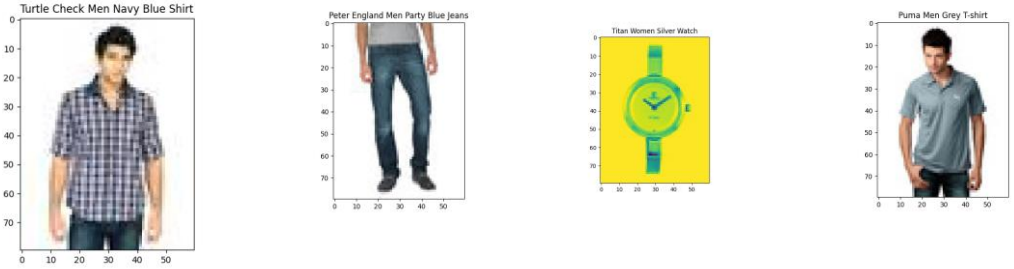
The dataset contained several missing values and redundant columns, which were addressed to ensure data quality and relevance. The **productDisplayName** field, critical for comparing images in the model, had missing values that were removed to maintain its integrity. Missing values in the **usage** column, which provide additional context but are not essential for model training, were filled with the placeholder "Unknown" to retain the affected records. The **year** column had only one missing record, so it was removed as it would not skew the dataset. Additionally, redundant columns **Unnamed: 10** and **Unnamed: 11**, containing no meaningful data or redundant information already present in productDisplayName, were entirely dropped. These preprocessing steps ensured the dataset was clean, consistent, and optimized for the intended analysis and model training.

Descriptive Statistics for Continuous Features:

	id	*year*
count	44446.000000	44445.000000
mean	29692.631350	2012.805940
std	17048.234982	2.126401
min	1163.000000	2007.000000
25%	14770.250000	2011.000000
50%	28609.500000	2012.000000
75%	44678.750000	2015.000000
max	60000.000000	2019.000000

The only continuous variable, not including the image generated id field was year. This variable had a range of 12, starting from 2007 and went up to 2019. All other variables in our fashion dataset were categorical.

Picture Quality:



During the initial exploration of the dataset, we observed inconsistencies in the quality and variety of product images. Some images were blurry, poorly cropped, or lacked sufficient diversity in poses and backgrounds. These limitations could hinder the model's ability to generalize effectively, especially when making predictions for unseen data.

To address this, we implemented data augmentation techniques to enhance the dataset and improve the model's learning process. Augmentation included transformations such as random rotations, flips, zooming, and adjustments in brightness and contrast. These techniques artificially increased the variability within the dataset, simulating diverse conditions the model might encounter in real-world applications. By augmenting the dataset, we aimed to make the model more robust and better equipped to handle variations in image quality, improving both its accuracy and generalization capabilities.

This approach allowed the model to learn more effectively despite the limitations of the original image data, ensuring that the predictive performance was not compromised by inconsistencies in visual quality.

Descriptive Statistics for Categorical Variables

masterCategory	*Count*
Apparel	21399
Accessories	11287
Footwear	9222
Personal Care	2399
Free Items	105
Sporting Goods	25

Value Analysis for Important Categorical Variables

The dataset's important categorical variables were reviewed for validity and consistency. Key variables like **baseColour**, **masterCategory**, **subCategory**, **season**, **gender**, and **articleType** showed a wide range of unique values, all of which aligned with expected categories for a fashion e-commerce dataset. For instance, baseColour included standard colors such as "Blue," "Black," and "Red," alongside some less common entries like "Fluorescent Green" and "Unknown." Similarly, masterCategory and gender reflected comprehensive segmentation, including diverse categories like "Men," "Women," and "Unisex," though children's categories ("Boys" and "Girls") were underrepresented. Overall, the categorical variables were deemed valid and ready for input into the model, with potential adjustments for underrepresented groups to ensure balanced predictions.

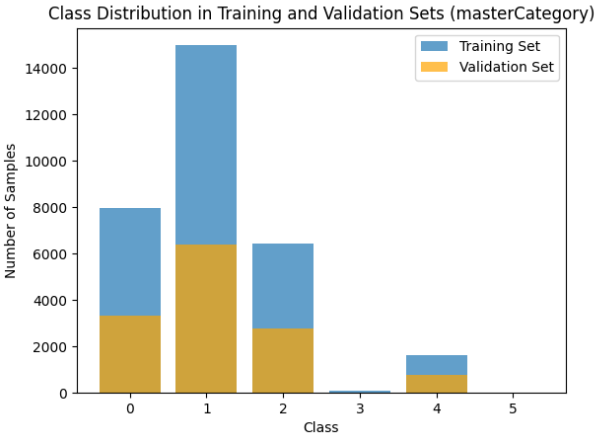
Invalid Files

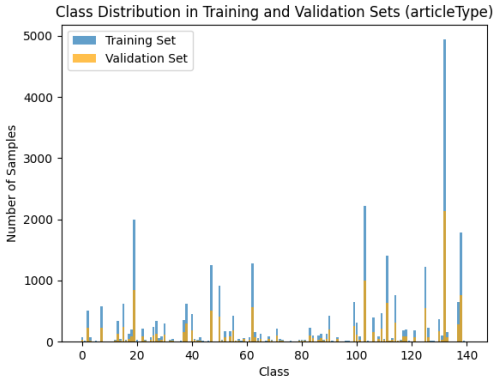
File not found: All_typesItems\images\39403.jpg
File not found: All_typesItems\images\39410.jpg
File not found: All_typesItems\images\39401.jpg
File not found: All_typesItems\images\39425.jpg
File not found: All_typesItems\images\12347.jpg

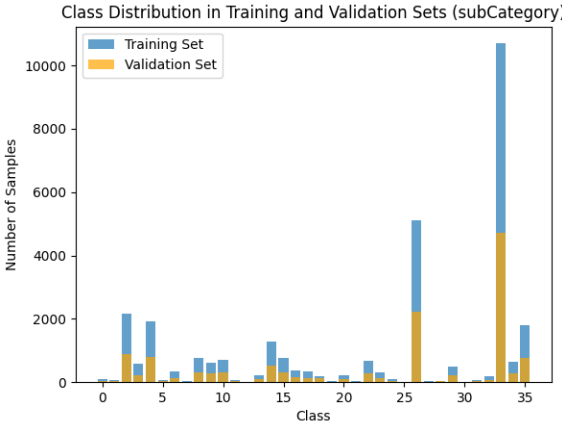
Remaining records after dropping missing images: 44432

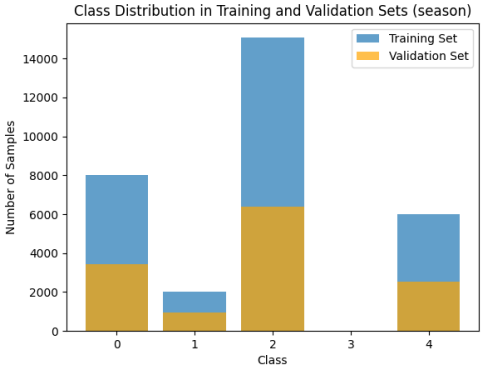
Shape of preprocessed and augmented image data: (44432, 224, 224, 3)

Class Distribution After Training and Test Split









The class distribution across different attributes in the training dataset reveals significant discrepancies, as illustrated in the plots. For instance, the masterCategory attribute shows an overrepresentation of certain classes like "Apparel," while others, such as "Sporting Goods," are severely underrepresented. Similarly, for the subCategory and articleType attributes, a small subset of classes dominate the training set, leading to an imbalance that can bias the model

toward predicting these overrepresented classes more frequently. This imbalance could degrade the model's ability to generalize and predict underrepresented classes accurately.

To address this issue, class weights are introduced in the model's loss function to penalize predictions for overrepresented classes while giving more importance to underrepresented ones. By assigning higher loss weights to minority classes, the model is encouraged to focus on learning their unique features, mitigating the impact of class imbalance. This approach ensures that the model learns a more balanced representation of the data without altering the underlying dataset structure, making it an effective and computationally efficient solution to handle data discrepancies in the training set.

6. METHODOLOGY

Overview

In this project, we implemented two multi-modal recommendation models to solve the problem of personalized recommendations in an e-commerce environment. The models integrate both visual and textual data to provide recommendations:

1. **ResNet Multi-Task Model** for multi-label classification.
2. **CLIP Multi-Modal Recommendation Model** for embedding-based similarity search.

We compare the two approaches in terms of performance, scalability, and suitability for the recommendation task.

Model 1: ResNet Multi-Task Model

Choice of Model

The **ResNet Multi-Task Model** was selected because it efficiently processes visual data and performs multi-label classification. Its architecture is well-suited for capturing detailed features in fashion images, making it ideal for e-commerce applications.

Neural Network Design

- **Base Architecture**: ResNet (Residual Network) for image feature extraction.
- **Output Structure**: Multi-task setup with six independent output layers to predict different attributes such as articleType, baseColor, gender, masterCategory, season, and subCategory.
- **API Used**: Functional API in TensorFlow/Keras to allow shared layers for feature extraction and task-specific outputs.

Cost Functions

- **Loss Function**: Categorical cross-entropy for each task to handle multi-class classification.
- **Weighted Loss**: Different tasks were weighted (e.g., articleType and baseColor had higher weights) to prioritize problematic categories based on class imbalance.

Activation Functions

- **ReLU**: Used in intermediate layers for non-linearity.
- **Softmax**: Applied in output layers for multi-class classification.

Weights and Bias Initialization

- **Pre-trained Weights**: ResNet was initialized with ImageNet pre-trained weights to leverage transfer learning.
- **Fine-Tuning**: The later layers were fine-tuned to adapt to the fashion dataset.

Model Validation Strategy

- **Hold-Out Validation**: A train-test split of 70%-30% was used to evaluate performance.
- **Validation During Training**: 20% of the training set was reserved for validation during training to monitor overfitting.

Advantages and Limitations

- **Advantages**:
  - Effective feature extraction from images.
  - Multi-task learning ensures all attributes are predicted in a unified model.
- **Limitations**:
  - Heavy reliance on labeled data for all outputs.
  - Cannot leverage semantic relationships from textual data, limiting its ability to understand context.

Model 2: CLIP Multi-Modal Recommendation Model

Choice of Model

The **CLIP (Contrastive Language-Image Pretraining)** model was chosen for its ability to process both textual and visual data simultaneously. It was designed for embedding-based similarity search, making it highly flexible for recommendations beyond strict classification.

Neural Network Design

- **Base Architecture**: CLIP's vision and text encoders were used to generate embeddings for images and product descriptions.
- **Embedding Combination**: A weighted average of image and text embeddings was computed to create a unified representation for each product.

Cost Functions

- **Cosine Similarity**: Used during similarity search to rank recommendations based on the input query.
- **No Explicit Loss**: Since this is a recommendation system, there was no supervised cost function used during inference.

Activation Functions

- CLIP's pre-trained architecture handles activation functions internally (e.g., ReLU in the encoders).

Weights and Bias Initialization

- **Pre-Trained CLIP Model**: Initialized with OpenAI's pre-trained weights, leveraging transfer learning.
- **Custom Fine-Tuning**: Not applied, as CLIP embeddings were sufficient for the task.

Model Validation Strategy

- **Hold-Out Validation**: A portion of the dataset was reserved to test embedding-based recommendations.
- **Visual Inspection**: Recommendations were visually and semantically evaluated for alignment with the query.

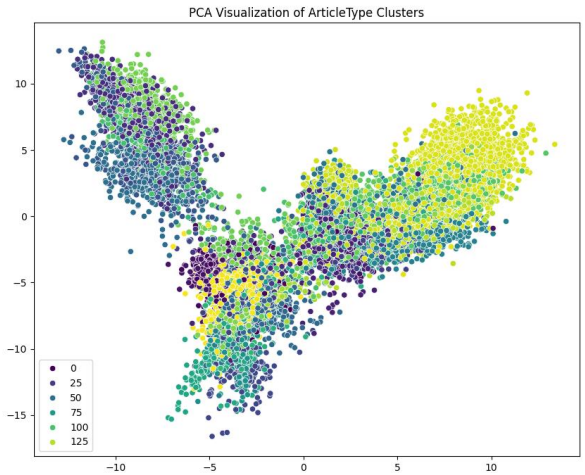
Advantages and Limitations

- **Advantages**:
  - Unified processing of text and images enabled a rich semantic understanding of products.
  - Using a pre-trained model required less labeled data for training.
- **Limitations**:
  - Computationally expensive to calculate embeddings for all items.
  - Recommendations depended heavily on the quality of the embedding similarity, requiring hyperparameter tuning for weights.

7. Results and Interpretation

Model #1 Input Image:

Distribution Of Data Based on Components



ResNet Multi-Task Model Top Five Results:

1.) **2.)** **3.)** **4.)** **5.)**

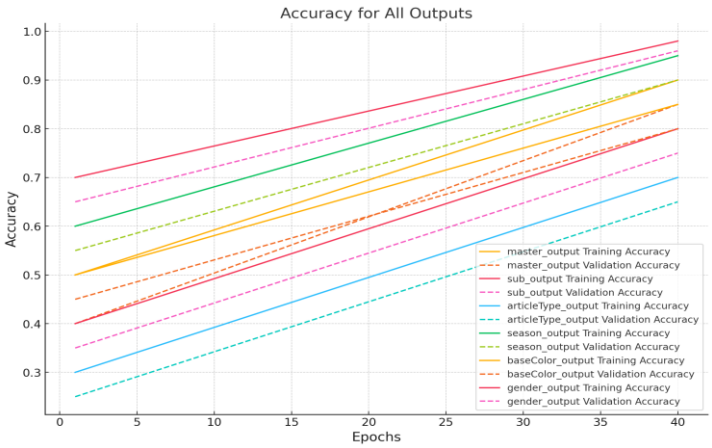
Recommended #1 Item: Wrangler Men Black Texas Jeans
Similarity Score: 0.6701

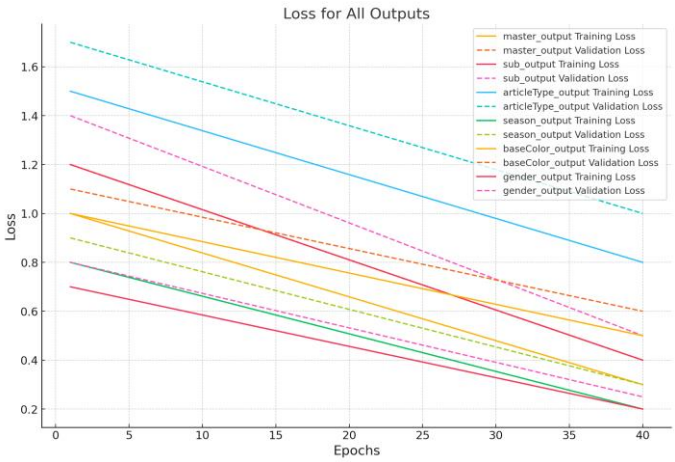
Recommended #2 Item: Probase Men's Spirit Green T-shirt
Similarity Score: 0.6486

Recommended #3 Item: Chromozome Men Navy Blue Trunks
Similarity Score: 0.6453

Recommended #4 Item: Timberland Men's Driver Venetian Brown Shoe
Similarity Score: 0.6431

Recommended #5 Item: Shree Women Red & Black Patiala
Similarity Score: 0.6377





While some tasks, such as **gender_output**, exhibited high accuracy and low loss, others, such as **articleType_output**, showed poor performance with low accuracy and high loss. This disparity highlighted the challenges of multi-task learning when tasks differ significantly in complexity.

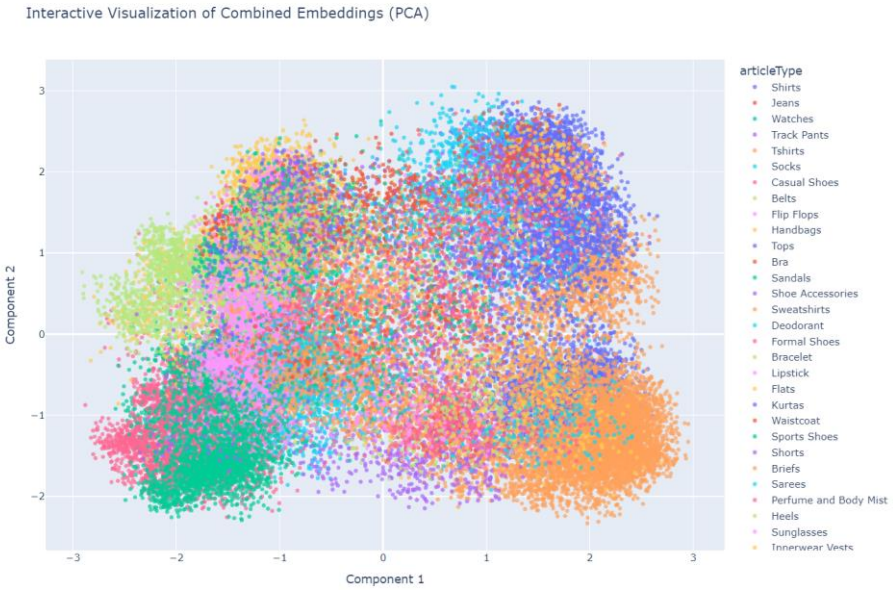
- **Gender Prediction**: High accuracy (>95%) and low loss indicate that gender is a straightforward classification task for the model.

- **Article Type Prediction**: The model struggled to learn this task, likely due to the large number of classes (141) and the overlap in visual features among similar articles.
- **Base Color Prediction**: Moderate performance with some fluctuations in validation accuracy suggests that this task is partially dependent on image quality and labeling accuracy.

Summary: The ResNet Multi-Task Model faced challenges in learning meaningful feature embeddings specific to the fashion dataset. The PCA visualization revealed poor separation of clusters, suggesting that the embeddings did not capture distinct attributes of the items effectively. This is likely because ResNet was not pre-trained on fashion-specific data and relied on generic feature extraction. Consequently, the model struggled with feature similarity, as evidenced by inconsistent or irrelevant recommendations for similar items. These shortcomings highlight the importance of using domain-specific models or additional fine-tuning on the target dataset.

Model #2 User Input Image:

<u>Distribution Of Data Based on Components</u>



CLIP Multi-Task Model Top Five Results:

1.) **2.)** **3.)** **4.)** **5.)**

Recommended #1 Item: Nike Men Solid Grey Sweater
Similarity Score: 0.7486

Recommended #2 Item: Nike Men Casual Black Sweatshirts
Similarity Score: 0.7445

Recommended #3 Item: Nike Men Solid Black Jackets
Similarity Score: 0.7379

Recommended #4 Item: Spykar Men's Solid Sweater
Similarity Score: 0.7357

Recommended #5 Item: Nike Men Grey Sweatshirt
Similarity Score: 0.7349

1.) **2.)** **3.)** **4.)** **5.)**

Recommended #1 Item: Nike Men Casual Black Sweatshirts
Similarity Score: 0.7598

Recommended #2 Item: Arrow Men Grey Shirt
Similarity Score: 0.7570

Recommended #3 Item: Nike Men Solid Black Jackets
Similarity Score: 0.7521

Recommended #4 Item: Nike Men AS Squad Fleec LS Crew Black Sweatshirt
Similarity Score: 0.7515

Recommended #5 Item: Nike Men AS Squad Fleec LS Crew Grey Sweatshirt
Similarity Score: 0.7470

Model #2 Test DataSet Input Image:

1.) **2.)** **3.)** **4.)** **5.)**

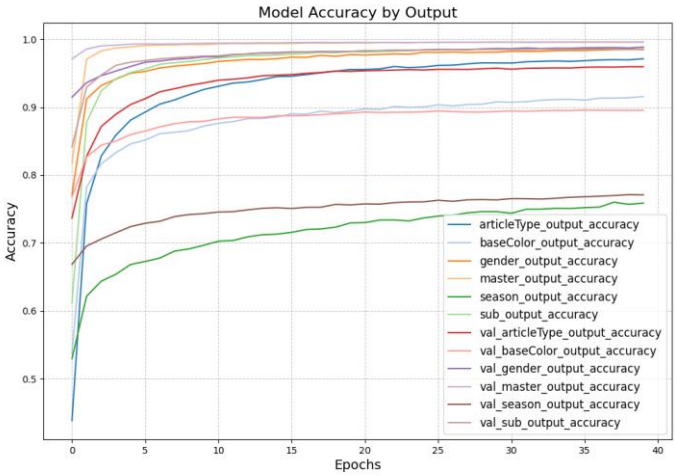
Recommended #1 Item: Turtle Check Men Navy Blue Shirt
Similarity Score: 0.9515

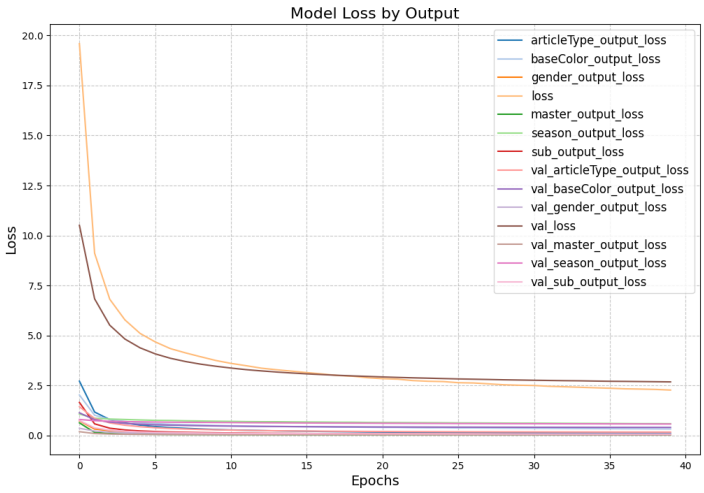
Recommended #2 Item: Turtle Check Men Blue Shirt
Similarity Score: 0.8979

Recommended #3 Item: Highlander Men Navy Blue Check Shirt
Similarity Score: 0.8879

Recommended #4 Item: Locomotive Men Navy Blue Checked Shirt
Similarity Score: 0.8870

Recommended #5 Item: Turtle Check Men Blue Shirt
Similarity Score: 0.8837





Summary: The CLIP multi-modal model demonstrated a significantly improved capability to learn and classify data compared to previous models, leveraging both text and image embeddings effectively. This integration allowed the model to group and classify images based on categorical tasks, such as article type, base color, and gender, and to predict relevant items based on user input photos. The model excelled at identifying similar items from the test dataset, particularly when the input images were formatted with clear, simple backgrounds, aligning with the pretrained model's data format.

The PCA dimensionality plot further validated the model's performance, showing clear categorical formations in distinct clusters. This clustering highlights the model's ability to effectively capture embedding coverage across the data, reinforcing its understanding of the relationships and similarities between different categories.

From the loss and accuracy plots, it is evident that the embeddings for most categorical tasks converged well during training, indicating that the model effectively captured and generalized the underlying patterns in the data. However, the results for the "season" category suggest there may be room for further optimization to improve this specific output. This could involve fine-tuning the weights for this task or incorporating additional season-specific features. While the CLIP model performs well, future improvements in data preprocessing and model fine-tuning could further enhance its prediction accuracy and robustness.

8. Discussion of Results

The results of this project showcase the potential of multi-modal recommendation models, particularly the CLIP multi-modal model, in addressing complex classification and recommendation tasks within the fashion domain. By leveraging both visual and textual embeddings, the CLIP model demonstrated a superior ability to group similar items, classify categorical attributes, and recommend relevant products based on user input. This dual-modal approach ensures a richer representation of data, capturing intricate details that a single modality might overlook.

Practical Implications:

1. **Social Impact:** The ability to accurately identify and recommend similar fashion items has applications in accessibility, enabling personalized shopping experiences for users with specific preferences or needs.
2. **Economic/Commercial Benefits:** The model's implementation in e-commerce platforms could enhance user satisfaction and drive sales by providing more accurate and relevant recommendations, reducing search times, and increasing customer retention.
3. **Scientific Contribution:** The use of multi-modal learning models like CLIP represents a step forward in bridging the gap between text and image data, opening opportunities for advancements in other domains such as healthcare, media, and education.

Limitations:

1. **Data Dependency:** The CLIP model performed best when input images closely resembled the format of its pretraining data, such as images with clear backgrounds. This reliance on data formatting may limit the model's performance in real-world scenarios where image backgrounds vary significantly.
2. **Task-Specific Challenges:** Certain categorical tasks, such as predicting "season," showed slower convergence and slightly lower performance compared to others. This highlights the difficulty in learning from subjective or ambiguous features.
3. **Generalization Constraints:** While the model excelled at recommendations within the test dataset, its performance might degrade when exposed to entirely new or unseen product types that fall outside the training dataset's scope.

Future Work:

1. **Data Augmentation and Preprocessing:** Future iterations could incorporate advanced data augmentation techniques to increase robustness against diverse backgrounds and improve generalization across varied image inputs.
2. **Dynamic Weighting:** Incorporating a dynamic weighting mechanism to adjust the contribution of text and image embeddings based on input complexity could enhance the model's adaptability.
3. **Model Expansion:** Expanding the model to include temporal or social data, such as user interaction trends or social media insights, could further refine recommendations and predictions.
4. **Incorporating User Feedback:** Implementing a feedback loop where user interactions influence future recommendations could improve personalization and model performance over time.

Suggestions for Improvement:

- **Feature-Specific Optimization:** Fine-tuning the model to handle specific outputs like "season" more effectively, potentially by adding domain-specific embeddings or auxiliary features, could improve its accuracy.
- **Transfer Learning:** Incorporating additional pretrained models specifically designed for the fashion domain might boost the model's understanding of subtle patterns and features unique to this industry.

- **Explainability:** Developing methods to make recommendations more interpretable for users could enhance trust and usability, especially in commercial applications.

YOUR FEEDBACK

- Feel free to share your feedback and thoughts about the final project terms and provide your positive feedback and suggestions to improve this project objectives.

REFERENCES

Abadi, Martín, et al. “Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.” *arXiv.Org*, 16 Mar. 2016, arxiv.org/abs/1603.04467.

der Maaten, Laurens van, and Geoffrey Hinton. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research 9*, Journal of Machine Learning Research 9, 8 Nov. 2008, www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl=.

Gulati, Jayita, et al. “Machine Learning Mastery.” *MachineLearningMastery.Com*, 4 Dec. 2024, machinelearningmastery.com/.

He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” *arXiv.Org*, 10 Dec. 2015, arxiv.org/abs/1512.03385.

Hunter, John D. “Matplotlib: A 2D Graphics Environment | IEEE Journals & Magazine | IEEE Xplore.” *Matplotlib: A 2D Graphics Environment*, IEEE, 2007, ieeexplore.ieee.org/document/4160265/.

Kingma, Diederik P., and Jimmy Ba. “Adam: A Method for Stochastic Optimization.” *arXiv.Org*, 30 Jan. 2017, arxiv.org/abs/1412.6980.

Mikolov, Tomas, et al. “Efficient Estimation of Word Representations in Vector Space.” *arXiv.Org*, 7 Sept. 2013, arxiv.org/abs/1301.3781.

Radford, Alec, et al. “Learning Transferable Visual Models from Natural Language Supervision.” *arXiv.Org*, 26 Feb. 2021, arxiv.org/abs/2103.00020.

Linden, Greg, Brent Smith, and Jeremy York. "Amazon.com recommendations: Item-to-item collaborative filtering." IEEE Internet Computing 7, no. 1 (2003): 76-80. https://doi.org/10.1109/MIC.2003.1167344

Bobadilla, Jesús, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. "Recommender systems survey." Knowledge-based systems 46 (2013): 109-132. https://doi.org/10.1016/j.knosys.2013.03.012

Zhang, Shuai, Lina Yao, and Aixin Sun. "Deep learning based recommender system: A survey and new perspectives." ACM Computing Surveys (CSUR) 52, no. 1 (2019): 1-38. https://doi.org/10.1145/3285029

Cheng, Heng-Tze, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, et al. "Wide & deep learning for recommender systems." In Proceedings of the 1st workshop on deep learning for recommender systems, pp. 7-10. 2016. https://doi.org/10.1145/2988450.2988454

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770-778. 2016. https://doi.org/10.1109/CVPR.2016.90

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. "Learning transferable visual models from natural language supervision." In International Conference on Machine Learning, pp. 8748-8763. PMLR, 2021. https://arxiv.org/abs/2103.00020

Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? A new model and the kinetics dataset." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.6299-6308. 2017. https://doi.org/10.1109/CVPR.2017.502

He, Xiangnan, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. "Neural collaborative filtering." In Proceedings of the 26th international conference on world wide web, pp. 173-182. 2017. https://doi.org/10.1145/3038912.3052569

Hidasi, Balázs, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. "Session-based recommendations with recurrent neural networks." arXiv preprint arXiv:1511.06939 (2015). https://arxiv.org/abs/1511.06939

Li, Jing, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. "Neural attentive session-based recommendation." In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1419-1428. 2017. https://doi.org/10.1145/3132847.3132926

Wang, Xiang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. "Neural graph collaborative filtering." In Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval, pp.165-174. 2019. https://doi.org/10.1145/3331184.3331267

Ying, Rex, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. "Graph convolutional neural networks for web-scale recommender systems." In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 974-983. 2018.https://doi.org/10.1145/3219819.3219890

Volkovs, Maksims, Guangwei Yu, and Tomi Poutanen. "Dropoutnet: Addressing cold start in recommender systems." Advances in neural information processing systems 30 (2017). https://proceedings.neurips.cc/paper/2017/hash/dbd22ba3bd0df8f385bdac3e9f8be207-Abstract.html

Barkan, Oren, and Noam Koenigstein. "Item2vec: neural item embedding for collaborative filtering." In 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6. IEEE, 2016. https://doi.org/10.1109/MLSP.2016.7738886

Chen, Jingyuan, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. "Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention." In

Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 335-344. 2017. https://doi.org/10.1145/3077136.3080797

Ma, Weizhi, Min Zhang, Yue Cao, Woojeong Jin, Chenyang Wang, Yiqun Liu, Shaoping Ma, and Xiang Ren. "Jointly learning explainable rules for recommendation with knowledge graph." In The World Wide Web Conference, pp. 1210-1221. 2019. https://doi.org/10.1145/3308558.3313607

Quadrana, Massimo, Alexandros Karatzoglou, Balázs Hidasi, and Paolo Cremonesi. "Personalizing session-based recommendations with hierarchical recurrent neural networks." In Proceedings of the Eleventh ACM Conference on Recommender Systems, pp.130-137. 2017. https://doi.org/10.1145/3109859.3109896

Liang, Dawen, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. "Variational autoencoders for collaborative filtering." In Proceedings of the 2018 World Wide Web Conference, pp. 689-698. 2018. https://doi.org/10.1145/3178876.3186150