Multimodal Explainable Recommender

Aasmaan Gupta (IMT2021006)
Trupal Patel (IMT2021056)
Kevin Adesara (IMT2021070)
Anant Ojha (IMT2021102)

Dataset

- We utilised the Amazon Fashion Dataset to develop our Fashion Recommendation System.
- `cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/`
- It contains product reviews and metadata from Amazon's Fashion category. It includes information about products, reviewers, ratings, and review text.

Dataset Description

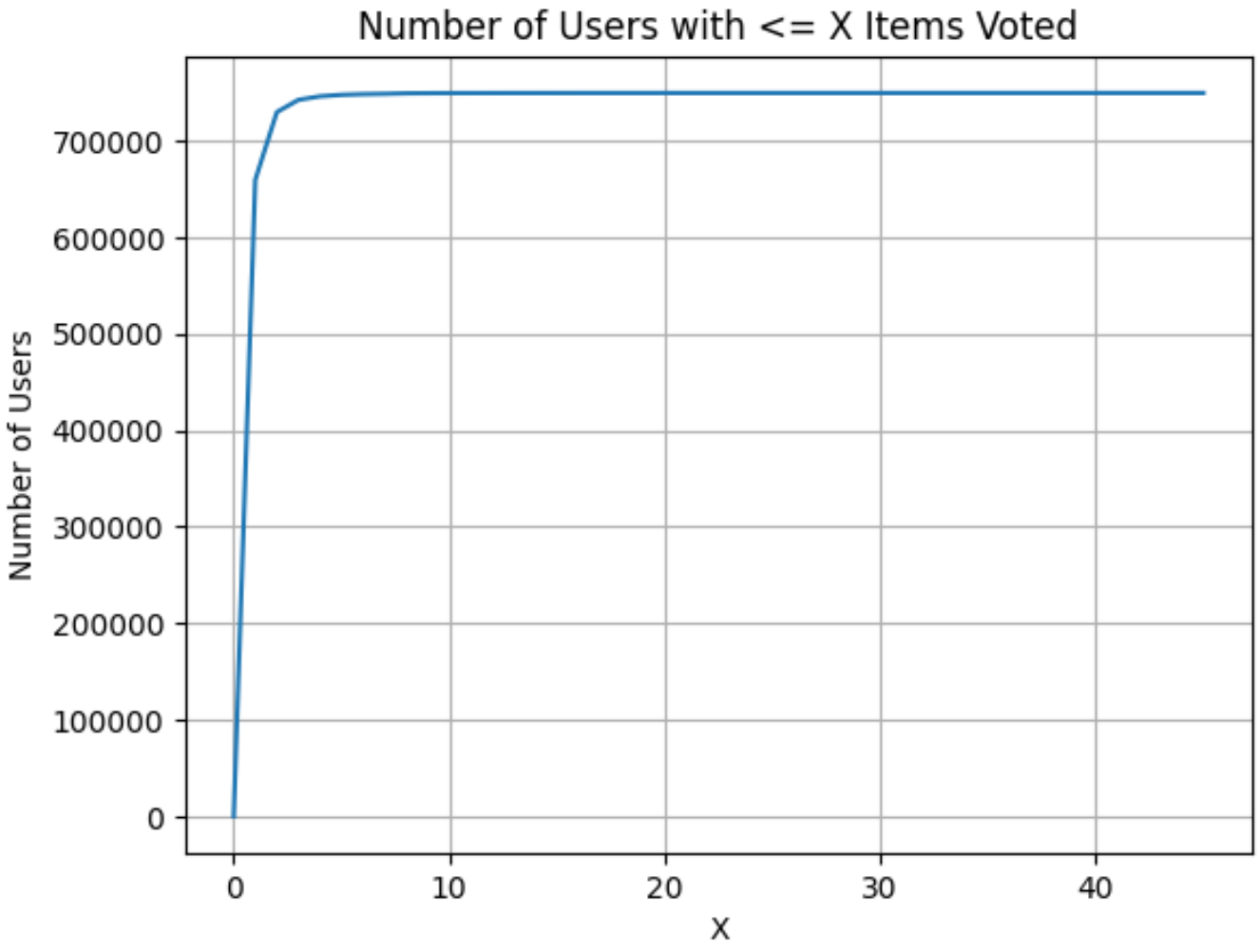
Review Data:

- Contains 883,636 reviews from various reviewers.
- Includes information such as reviewer ID, product ID, review text, rating, summary, and review time.

Metadata:

- Provides additional details about the products.
- Includes product ID, title, price, brand, and categories.

Dataset Analysis



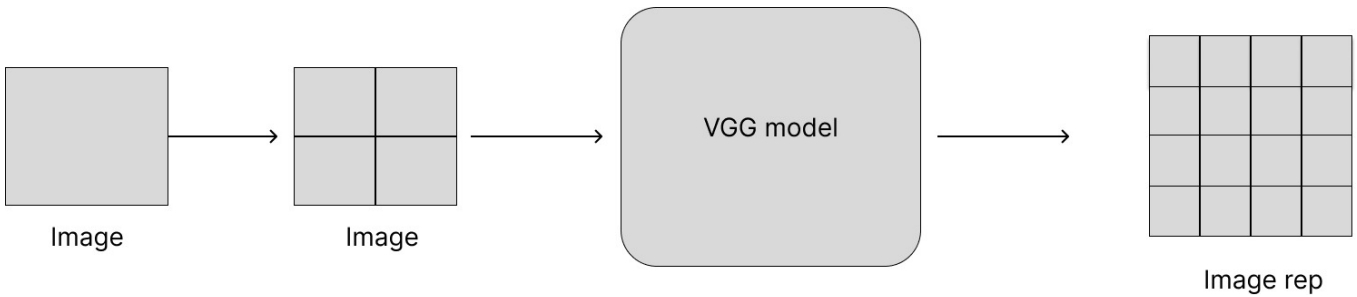
Data Preprocessing

- We only included reviews with no null values.
- Due to the large size of the dataset, we opted to use a subset of the data.

User ID	Item ID	ImageURL	Detail	Rating	Review

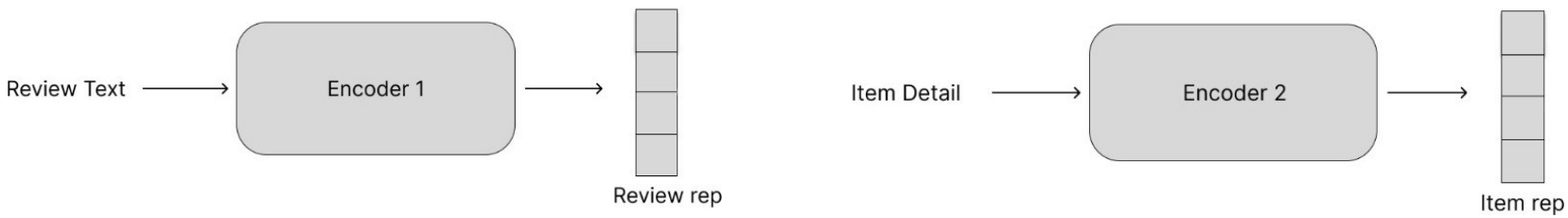
Data Preprocessing: Image Representations

- We utilized VGG to extract image representations.
- We considered 2x2 windows to obtain 4 vectors per image.



Data Preprocessing: Vector Representations

- We utilized encoders to generate vector representations for review text and item descriptions.



Introduction to Concepts

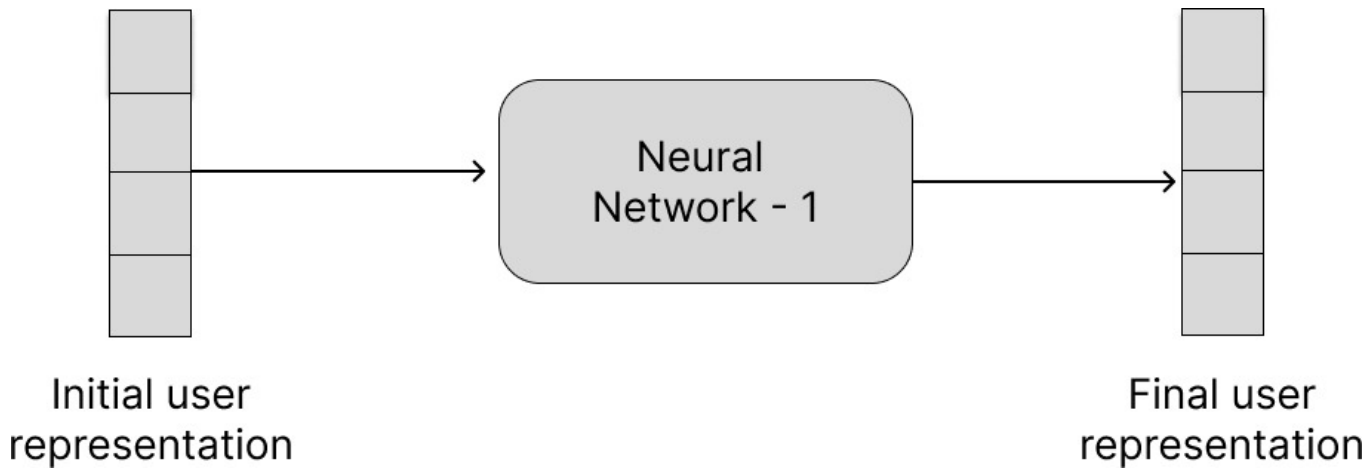
- We utilize neural networks to build a Multimodal Explainable Fashion Recommendation System.
- Our system involves training multiple neural networks to perform specific tasks:
  - **First Neural Network:** Obtains the final user representation.
  - **Second Neural Network:** Calculates the alphas, coefficients used to combine final image representations.
  - **Third Neural Network:** Predicts ratings using weighted image, final user representation, and item representation.

Core Idea

- Our architecture leverages both text and image modalities to effectively capture user preferences
- It intelligently assigns weights to various regions of an image based on the user's interests.
- By enhancing feature representations of images and combining them with user and textual item features, our model predicts relevant items.
- Additionally, our model provides explanations for its recommendations by highlighting areas of interest in recommended images.

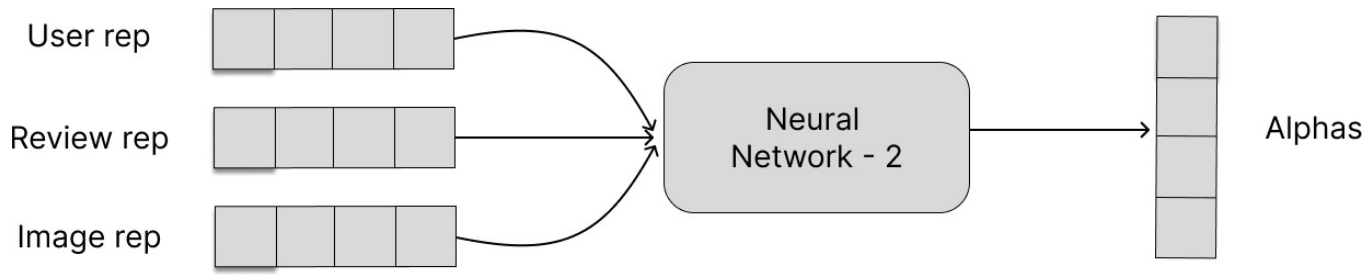
Architecture

- We train a neural network to obtain the final user representation.



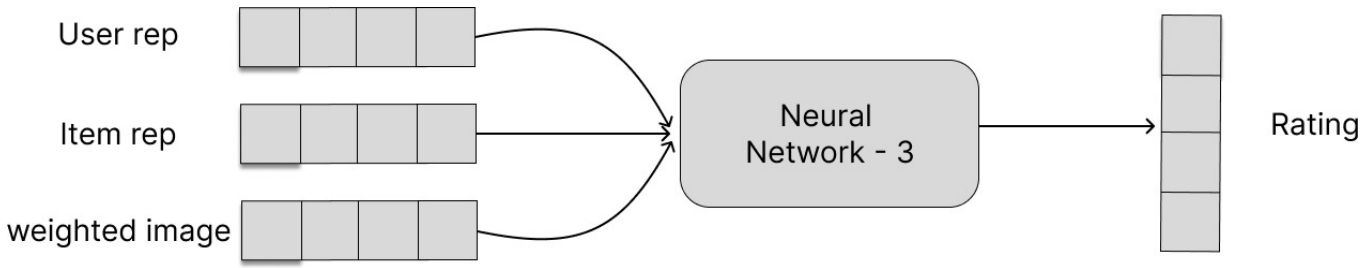
Architecture

- Another neural network is trained to calculate the alphas, which are the coefficients used to combine the final image representations.

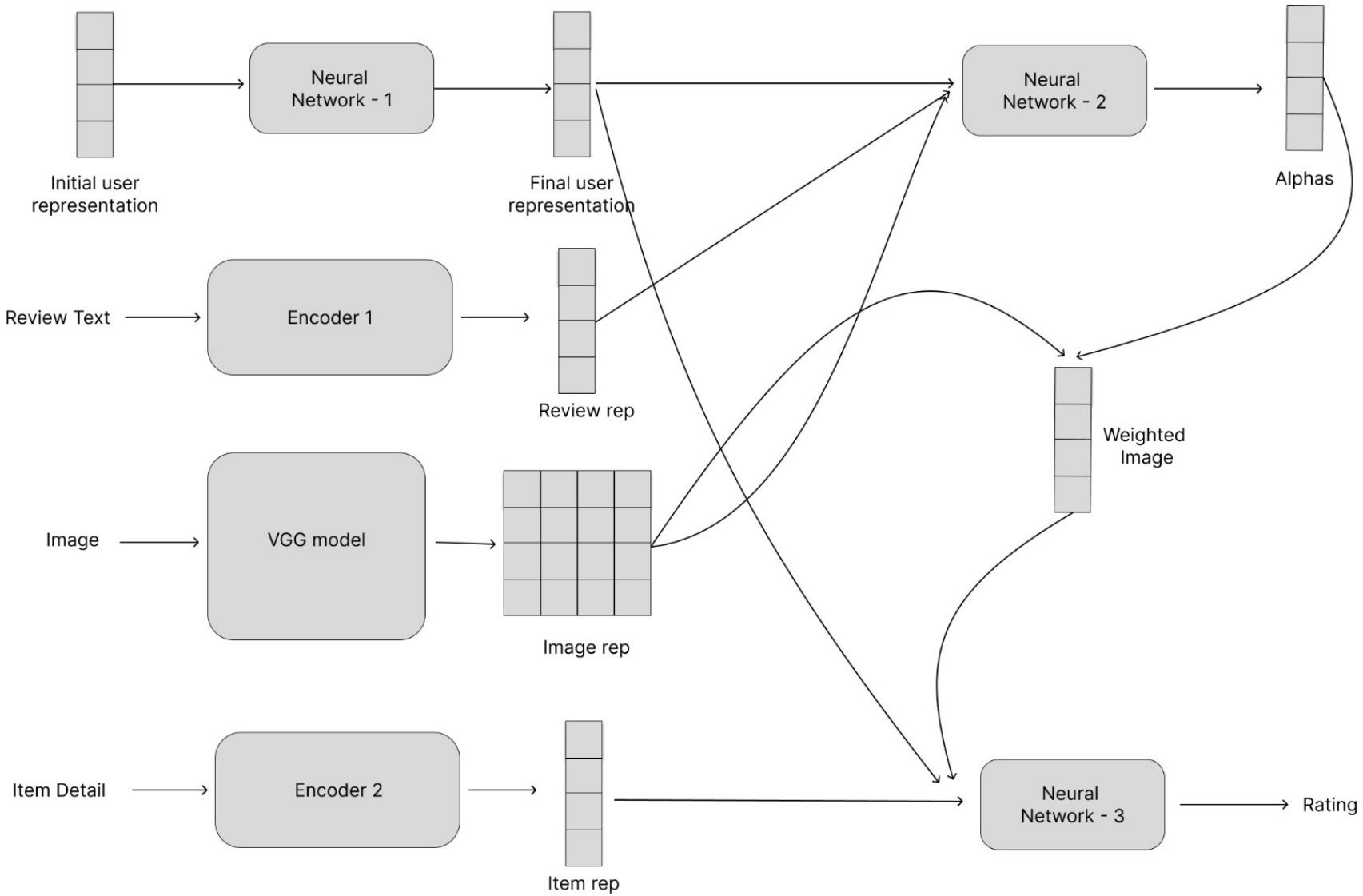


Architecture

- We feed the weighted image, final user representation, and item representation to a third neural network to predict ratings..



Everything in Action



Recommendation

- To generate recommendations, we take input text from the user specifying their preferences.
- This input text is fed into our neural network to obtain ratings for all the products in the dataset.
- From the top-rated products, we randomly select 5 items and display their pictures to the user.

Limitations

- Our encoder layers and representation sizes are minimal due to computational constraints.
- We are training our models on a dataset with only about ten thousand data entries.
- Image regions are divided into only four regions instead of more regions due to computational limitations.

Novelty

- We employ a multimodal approach within our attention mechanism. By extracting user preferences from reviews or input text, we dynamically assign weights to different areas of the image. This innovative method effectively translates textual user choices into the visual domain, enhancing the model's ability to accurately capture user preferences across modalities.
- Our model possesses the unique capability to provide explanations for its recommendations. By highlighting specific areas of interest within recommended images, based on the model's perception of potential regions of interest.