




## Article

# A Multimodal Recommender System Using Deep Learning Techniques Combining Review Texts and Images

Euiju Jeong <sup>1,†</sup>, Xinzhe Li <sup>1,†</sup> , Angela (Eunyoung) Kwon <sup>2</sup>, Seonu Park <sup>1</sup> , Qinglong Li <sup>1</sup>  and Jaekyeong Kim <sup>1,3,\*</sup>

<sup>1</sup> Department of Big Data Analytics, Kyung Hee University, 26, Kyungheedaero-ro, Dongdaemun-gu, Seoul 02447, Republic of Korea; euiju1011@khu.ac.kr (E.J.); lixz@khu.ac.kr (X.L.); sunu0087@khu.ac.kr (S.P.); leecy@khu.ac.kr (Q.L.)

<sup>2</sup> Sauder School of Business, University of British Columbia, 2053 Main Mall, Vancouver, BC V6T 1Z2, Canada; angela.kwon@sauder.ubc.ca

<sup>3</sup> School of Management, Kyung Hee University, 26, Kyungheedaero-ro, Dongdaemun-gu, Seoul 02447, Republic of Korea

\* Correspondence: jaek@khu.ac.kr; Tel.: +82-2-961-9355

† These authors contributed equally to this study.

**Abstract:** Online reviews that consist of texts and images are an essential source of information for alleviating data sparsity in recommender system studies. Although texts and images provide different types of information, they can provide complementary or substitutive advantages. However, most studies are limited in introducing the complementary effect between texts and images in the recommender systems. Specifically, they have overlooked the informational value of images and proposed recommender systems solely based on textual representations. To address this research gap, this study proposes a novel recommender model that captures the dependence between texts and images. This study uses the RoBERTa and VGG-16 models to extract textual and visual information from online reviews and applies a co-attention mechanism to capture the complementarity between the two modalities. Extensive experiments were conducted using Amazon datasets, confirming the superiority of the proposed model. Our findings suggest that the complementarity of texts and images is crucial for enhancing recommendation accuracy and performance.

**Keywords:** recommender system; multimodal review; deep learning; co-attention mechanism; complementarity



**Citation:** Jeong, E.; Li, X.; Kwon, A.; Park, S.; Li, Q.; Kim, J. A Multimodal Recommender System Using Deep Learning Techniques Combining Review Texts and Images. *Appl. Sci.* **2024**, *14*, 9206. <https://doi.org/10.3390/app14209206>

Academic Editors: João M. F. Rodrigues and Eleonora Iotti

Received: 29 August 2024

Revised: 22 September 2024

Accepted: 8 October 2024

Published: 10 October 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

With the development of Information and Communication Technology (ICT) and the widespread use of the Internet, the rapidly growing e-commerce market allows users to easily search for and purchase a variety of newly released products [1]. Despite these advantages, users face information overload problems when selecting items that suit their preferences [2,3]. This phenomenon has become a reason for e-commerce platforms to recognize the importance of recommender systems, which take into account user preference and recommend items that are likely to be purchased based on one's need [4]. The mechanism behind recommender systems is based on the analysis of past user behavior for the recommendations [5]. Therefore, users can reduce their information search costs, and companies can secure their competitive advantages for customer management and sales improvement.

Collaborative Filtering (CF) is one of the typical recommender systems that has been widely used [6]. However, since it solely relies on users' past behavior, failing to fully address users' behavioral motivations leads to a data sparsity problem [7]. To solve this issue, previous studies have used auxiliary information such as online reviews, which contain information related to user and item features [2,8]. For example, Zheng, Noroozi and Yu [6] used a Convolutional Neural Network (CNN) on review sets of users and items

for extracting textual representations for recommendations. Liu, Chen and Chang [8] used Bidirectional Encoder Representations from Transformers (BERT) and Robustly optimized BERT approach (RoBERTa) models as the encoder to extract textual information from review texts. The experimental results showed such methods can significantly enhance recommendation performance, as they incorporate more diverse information from the review texts. Although mitigating the issue of data sparsity and improving recommendation performance, these review-based recommender system studies have a limitation in that they fail to capture user preferences for visual representations [9].

Nowadays, most online reviews are composed of multimodal contents including both images and texts [10]. Images with visual information can convey detailed aspects of items and users' preferences in an intuitive manner that review texts may not be able to deliver [11]. From the sentence "I love this design" in the review text shown in Figure 1, it is evident that the user appreciates a specific aspect of the phone case design. However, relying solely on the text, it is difficult to determine which specific part of the design the user appreciates. In other words, images can directly deliver user preferences that are not expressed in the text [9]. Therefore, texts as textual information and images as visual information in multimodal reviews provide different types of information that can be complementary or substitutive [11]. Meanwhile, some studies confirm that fusing these two types of information enhances the effectiveness of the recommendations [10]. Therefore, it is necessary to propose a recommender system that can capture complementary and substitutive effects between texts and images.



**Figure 1.** Example of an online review of Amazon.

For the fusion strategy, most studies used multiplication or concatenation operations, which do not fully reflect the interactions between different modalities [12]. To achieve the proposed idea, this study applies the advanced fusion method based on attention mechanism to capture the complementarity between texts and images. Specifically, this study leverages the advantages of co-attention to model the complementarity between review texts and images by focusing on aligned features. Therefore, this study proposes a novel recommender system model called CAMRec (Co-Attention based Multimodal Recommender system), which effectively integrates the complementary effects of texts and images into the recommendation tasks. First, we use the RoBERTa model to extract textual information, which demonstrates excellent performance on various Natural Language Processing (NLP) tasks. Second, we use the VGG-16 model to extract visual information, which effectively recognizes complex patterns and features of images. Third, we apply a co-attention mechanism to obtain the joint representations of textual and visual information through dependencies between them. Finally, we integrate the overall features to predict user preference for a specific item. To evaluate the recommendation performance of the proposed CAMRec model, we use two datasets from Amazon. The experimental results confirmed that the proposed CAMRec outperforms various baseline models. The primary contributions of this study can be summarized as follows:

- This study proposed CAMRec, which reflects user preferences from textual and visual perspectives and provides recommendations based on the complementarity between the two modalities.
- This study explored the impact of users' perception of visual representations on performance in multimodal recommender systems. It offers valuable insights into how visual features reflect user preferences.
- This study conducted extensive experiments using real-world datasets from Amazon. The experimental results offer new directions for future research in the field of recommender systems.

The rest of this study is as follows: Section 2 addresses related works, and Section 3 describes our proposed CAMRec model. Section 4 outlines the experimental data, evaluation metrics, and baseline models used in this study. Section 5 shows the experimental results and discussions about them. Finally, Section 6 concludes this study.

## 2. Related Works

### 2.1. Review-Based Recommender System

CF, as a typical recommender system, provides recommendations based on users' past behaviors, such as ratings and browsing history [13,14]. However, it is limited in addressing users' behavioral motivations and data sparsity issues [15]. Therefore, review texts are the solution to alleviate such limitations as they contain information on users' specific preferences for items [4]. Many studies on recommender systems extract textual features from review texts and incorporate them into the recommendations. For example, Chen et al. [16] used user and item review sets with CNN and attention mechanisms to extract helpful information. They found that such a method can improve the quality of embeddings, thereby improving recommendation performance. Cao et al. [7] provided recommendations using Word2Vec and a CNN to analyze the semantic content of review texts. The proposed model enables more sophisticated and accurate predictions than simple text analysis. Liu, et al. [5] proposed a hybrid model, which reflects the relationship between ratings and reviews to enhance the quality of latent user and item representations. Liu, et al. [8] extracted review text features with a multi-embedding perspective, which embedded user reviews using BERT and RoBERTa models, respectively. They indicated that such a method can extract richer and deeper features than individual review texts to obtain enhanced textual representations.

Although these review-based recommender system studies have alleviated the data sparsity issue and improved recommendation performance, they solely focus on textual information [6,16]. In other words, they overlook the images in the recommendation tasks, which can provide visual information about the item and further represent users' preferences. Therefore, this study aims to propose a recommender system based on multimodal information to enhance the recommendation performance.

### 2.2. Multimodal Recommender System

With the surge in multimodal data, researchers have begun to pay closer attention to visual representations as a data source in recommender systems. Visual content, such as user-generated images, can show detailed features of items based on different perspectives. Images can not only convey users' preferences for items but also reduce the perceived uncertainty of textual information. Therefore, some studies have proposed multimodal recommender systems that fuse various data types to provide more accurate and personalized recommendations [9,17]. Compared to review-based recommender systems, multimodal recommender systems integrate multiple data sources to alleviate data sparsity issues further. He and McAuley [18] proposed a multimodal recommender system by integrating item visual features into Matrix Factorization (MF). The proposed model not only provides more accurate recommendations but also alleviates the cold start issues. Chen et al. [19] focused on the fashion domain and aimed to provide visual explanations for the recommendations. They applied the attention mechanism to extract users' segmented preferences on

images and integrate them with review texts. Meanwhile, Liu et al. [20] extracted individual feature representations from texts and images and incorporated the attention mechanism in computing the relative importance of each modality, which can assign weights for the modality based on the self-importance in the recommendation process.

Although these studies proposed multimodal recommender systems and commonly confirmed their effectiveness, the complementary or substitutive effects of multimodal data were overlooked. Indicated texts provide detailed information on various aspects of customer preferences, while images complement this by offering visual information. Xiao et al. [11] indicated that texts provide detailed information on various aspects of customer preferences, while images complement it by offering visual information. Meanwhile, Xu et al. [9] confirmed that considering the consistency between texts and images enhances the accuracy of recommendation performance. In the context of online reviews, texts and images can be complementary or substitutive, while previous studies are still limited in these effects. Therefore, this study proposes a novel model that introduces the complementarity of texts and images in the recommendation tasks.

### 2.3. Multimodal Fusion Techniques

Multimodal fusion is an essential topic in the deep learning domains, and the integration of image and text data is useful in various applications, such as sentiment analysis [21], visual question answering [22], review helpfulness prediction [23], and also recommender systems [20].

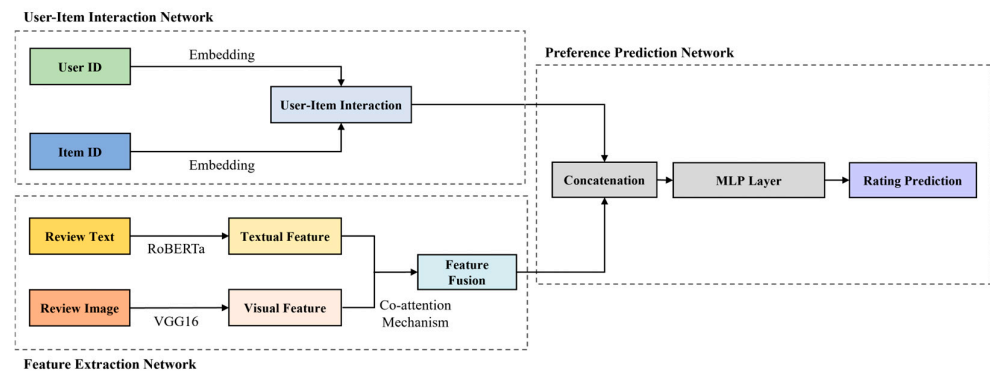
Early studies fused various multimodal data using relatively simple operations such as multiplication and concatenation. Although such methods are direct and easy to implement, they are limited in capturing complex interactions between different modalities [10]. To address this issue, recent studies indicated that the co-attention mechanism is the optimal solution for effectively integrating the interdependencies between modalities. For example, Lu et al. [24] showed that the co-attention mechanism can process information from both images and written questions. Laenen and Moens [25] compared and analyzed various fusion methods related to item features extracted from visual and textual information and confirmed that using the co-attention mechanism shows the best performance. Meanwhile, Liu et al. [12] modified self-attention into a co-attention between the source and target to generate mutually attended representations for the texts and labels. The proposed method focused on the relevant parts to enhance text classification performance. Ren et al. [10] also used the co-attention mechanism to obtain fused representations based on the dependencies between texts and images for review helpfulness prediction. In summary, the co-attention mechanism has been demonstrated to effectively capture the interactions between features at a deep level [12].

The goal of this study is to provide recommendations based on introducing the complementarity of review texts and images. Therefore, this study leverages the advantages of the co-attention mechanism to model the complementarity and integrate it into the recommender system.

## 3. Methodology

### 3.1. Problem Definition

The overall framework of the proposed CAMRec model is illustrated in Figure 2. The CAMRec model consists of User–Item Interaction Network, Feature Extraction Network, and Preference Prediction Network. Specifically, the User–Item Interaction Network captures the interactions between users and items. The Feature Extraction Network extracts textual and visual information from online reviews that can represent users' preferences and then fuse them. The Preference Prediction Network concatenates the overall features to predict the final rating a user will leave for a specific item.



**Figure 2.** Framework of the CAMRec Model.

Using information from both texts and images is a stepping stone to understanding users' preferences and behaviors for items in a comprehensive manner. Nevertheless, most studies have overlooked the informational value of images and have focused solely on texts. Meanwhile, although the fusion method is crucial for using multimodal data, previous studies are limited in capturing the dependencies and complementarity between modalities. Therefore, this study proposes the CAMRec model, which integrates the complementarity of textual and visual representations into the recommendations. The CAMRec model uses the RoBERTa and VGG-16 models to extract features from review texts and images, respectively. These models have been extensively validated and have been used to extract textual and visual features from review texts and images in previous studies [8,10]. Therefore, they ensure the reliability of our feature extraction process to support the proposed CAMRec model effectively. Moreover, the proposed CAMRec models the complementarity between texts and images using the co-attention mechanism, which is critical in this study. The co-attention mechanism addresses the task of integrating textual and visual features by dynamically aligning and fusing. It ensures that relevant features from each modality enhance the understanding of information from the other, thereby capturing their interdependencies. Finally, these joint representations of review texts and images enrich the features of the users and the items and are integrated with user–item interactions to perform the recommendation tasks.

Given  $T = (u, r, m, i, y)$  as a tuple composed of user  $u$ , item  $i$ , user's review text  $r$ , user's review image  $m$ , and rating  $y$ . The proposed model aims to train the prediction model  $A$  formalized as follows:

$$A(u, r, m, i, y; \theta) \rightarrow \hat{y}, \quad (1)$$

where  $\theta$  and  $\hat{y}$  indicate bias and predicted preference rating. During the training process, the model is trained to learn predicted preference rating  $\hat{y}_{u,i}$  based on the actual user preference rating  $y_{u,i}$ . The detailed description of each network of the CAMRec model follows in Section 3.2.

### 3.2. CAMRec Architecture

The CAMRec model proposed in this study is designed to effectively consider the complementarity of textual and visual features based on the co-attention mechanism. Figure 3 shows the entire framework of the proposed CAMRec model, and the detailed description for each network is as follows.

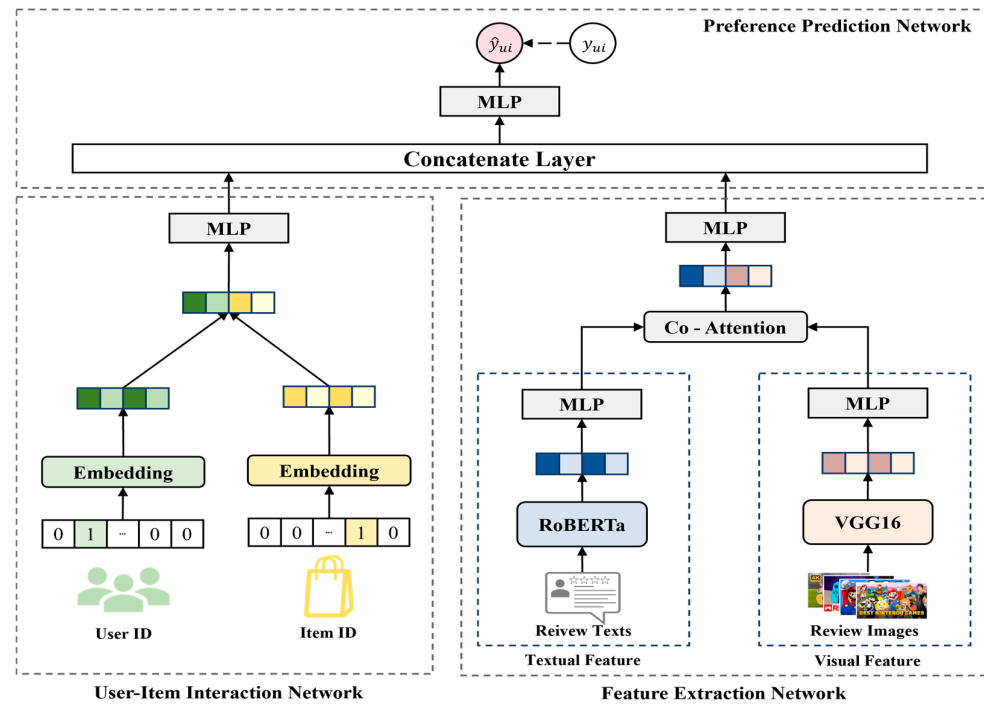


Figure 3. Architecture of the CAMRec Model.

### 3.2.1. User–Item Interaction Network

The User–Item Interaction Network aims to capture the complex interactions between the users and the items. First, to obtain the latent representation vectors  $p_u$  and  $q_i$  of the user  $u$  and the item  $i$ , we convert them to sparse vectors through one-hot encoding. Next, the converted sparse vectors are embedded for conversion into dense vectors as follows:

$$\begin{aligned} p_u &= P^T v_u^U, \\ q_i &= Q^T v_i^I, \end{aligned} \quad (2)$$

where  $v_u^U$  and  $v_i^I$  indicate the one-hot encodings for the user  $u$  and the item  $i$ , respectively.  $P \in \mathbb{R}^{m \times d}$  indicates the user embedding matrix and  $Q \in \mathbb{R}^{n \times d}$  indicates the item embedding matrix, while  $m$  and  $n$  indicate the number of users and items with  $d$  indicating dimensions of latent vectors. Then, the output of the user latent vector  $p_u$  and item latent vector  $q_i$  are concatenated as follows:

$$E = [p_u \oplus q_i], \quad (3)$$

where  $\oplus$  indicates the concatenation operation. Next, the concatenated vector  $E$  is fed into the Multi-Layer Perceptron (MLP) to learn the complex interactions through a nonlinear optimization process as follows:

$$\begin{aligned} E_1 &= \sigma(W_1^E E + b_1^E), \\ &\dots \\ E_M &= \sigma(W_M^E E_{M-1} + b_M^E), \end{aligned} \quad (4)$$

where  $W_M^E$ ,  $b_M^E$ , and  $\sigma$  indicate weight matrix, bias, and the Leaky Rectified Linear Unit (ReLU) activation function for each layer, respectively. The output of this network is vector  $E_M$ , which represents the interaction information between the user and the item.



### 3.2.2. Feature Extraction Network

The Feature Extraction Network aims to extract the textual and visual features from texts and images and fuse them to represent users' preferences.  $R_{u,i} = \{r_1, r_2, \dots, r_l\}$  represents a review text left by user  $u$  on item  $i$ , where  $l$  represents the length of a review text. Meanwhile,  $M_{u,i} = \{m_1, m_2, \dots, m_n\}$  represents the image accompanied by the review text, where  $n$  represents the number of images.

To extract the textual feature of review texts, we use the pre-trained RoBERTa model, which has demonstrated its excellent performance in the NLP field. The RoBERTa model can be seen as an extension of BERT, which was trained using a dynamic masking method, larger datasets, and additional training optimizations. The pre-trained RoBERTa base model consists of 12 transformer encoders, each with 12 attention heads, and each output token is represented as a 768-dimensional vector. Since the [CLS] token of RoBERTa contains detailed meanings behind the sentence [8], we use the 768-dimensional vector of the [CLS] token as the textual feature. Therefore, the textual feature  $O^{Text}$  can be defined as follows:

$$O^{Text} = \text{RoBERTa}(R_{u,i}), \quad (5)$$

To extract the visual feature of review images, we use the pre-trained VGG-16 model. VGG-16 is a deep CNN architecture widely used in image recognition and classification tasks. The VGG-16 model consists of 13 convolutional layers followed by three fully connected layers. Each convolutional layer employs  $3 \times 3$  filters to extract features from the input images. Before inputting images to the VGG-16 model, each image should be resized to  $224 \times 224$  pixels, converted to the RGB color space, and normalized by subtracting the mean of each color channel [26]. Since the model is used to perform image recognition, we decided to use the 4096-dimensional vector of the last convolutional layer as the visual feature. Considering the number of images varies with the reviews, we decided to extract the vectors of each image and average the value to represent the visual features for the review effectively. Therefore, the visual feature  $O^{Image}$  can be defined as follows:

$$O^{Image} = \frac{1}{n} \sum_{i=1}^n \text{VGG-16}(M_{u,i}), \quad (6)$$

where  $n$  represents the number of images. For both review text and image features, they are fed into the MLP to enhance nonlinearity and ensure the same dimension for each feature. Therefore, the final review text and image feature vectors can be represented as  $O_M^{Text}$  and  $O_M^{Image}$ .

The goal of this study is to complementarily fuse features based on the relative importance of each modality in the recommendation process. Therefore, this study leverages the advantages of the co-attention mechanism to model the complementarity between extracted final review text and image feature vectors. For a clear comparison, we recap the scaled dot product attention mechanism as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \quad (7)$$

where  $Q, K$ , and  $V$  indicate query, key, and value, respectively.  $\sqrt{d_k}$  indicates the square root of the dimension of the key vector  $d_k$ . The attention mechanism is typically implemented using a multi-head structure, where multiple attention heads operate in parallel to capture different aspects of the information. The formula of the multi-head attention mechanism can be defined as:

$$\text{MultiHead}(Q, K, V) = (\text{head}_1 \oplus \text{head}_2 \oplus \dots \oplus \text{head}_h) W^O, \quad (8)$$

where  $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ ,

where  $W_i^Q \in \mathbb{R}^{d_k \times d_p}$ ,  $W_i^K \in \mathbb{R}^{d_k \times d_p}$ , and  $W_i^V \in \mathbb{R}^{d_v \times d_p}$  indicate the weight matrices for the queries  $Q$ , keys  $K$ , and values  $V$  of the  $i$ -th head, respectively. Meanwhile,  $W^O \in \mathbb{R}^{hd_v \times d_p}$  indicates the weight matrix of the output layer, and  $h$  is the number of heads. The purpose of using multi-head is for capturing different aspects of information simultaneously by using multiple heads.

This study uses  $d_k = d_v = d_p/h = 64$ . The original multi-head self-attention mechanism [27] focuses on assigning weights based on computing the relative importance within a single modality, effectively capturing intra-modal relationships. However, the co-attention mechanism is proposed to capture inter-modal interactions between different modalities. This mechanism is consistent with our proposed idea to introduce and leverage the complementarity between review texts and images in recommendation systems.

For the self-attention mechanism,  $Q$ ,  $K$ , and  $V$  are all derived from the same modality. However, the co-attention mechanism uses the  $Q$  from the source modality while sourcing the  $K$  and  $V$  from the target modality. In other words, the co-attention mechanism ensures that the query vector from one modality attends to the relevant key and value vectors from the other modality. This allows the model to capture the complementary information between the two modalities by focusing on aligned features, thereby enabling an effective fusion of textual and visual features. Therefore, the process to capture the complementarity of textual and visual features can be defined as:

$$\begin{aligned} T_{Co} &= \text{MultiHead}(O_M^{\text{Text}}, O_M^{\text{Image}}, O_M^{\text{Image}}), \\ I_{Co} &= \text{MultiHead}(O_M^{\text{Image}}, O_M^{\text{Text}}, O_M^{\text{Text}}), \end{aligned} \quad (9)$$

where  $T_{Co}$  and  $I_{Co}$  indicate the text-attended representation and image-attended representation, respectively. Then, we follow the Transformer architecture [27] and apply residual connections and two independent Feed Forward Networks (FFN) through Layer Normalization (LN) to further enhance these co-attended features as follows:

$$\begin{aligned} T &= \text{LN}(\text{FFN}(T_{Co}) + O_M^{\text{Text}}), \\ I &= \text{LN}(\text{FFN}(I_{Co}) + O_M^{\text{Image}}), \end{aligned} \quad (10)$$

where  $T$  and  $I$  indicate the enhanced feature vectors. Next, we apply an element-wise operation to fuse the features as shown in Equation (11). This operation helps to capture the finer interactions between the co-attended textual and visual features by ensuring that only the corresponding features from both modalities contribute to the final fused representation.

$$F = T \odot I, \quad (11)$$

where  $F \in \mathbb{R}^{d_p}$  indicates the fused feature vector and  $\odot$  indicates the element-wise product operation. Specifically, the fused  $F$  feature vector combines the enriched textual and visual representations, allowing the model to leverage the complementary aspects of both modalities.

### 3.2.3. Preference Prediction Network

The Preference Prediction Network aims to predict the user's preference for a specific item using the output feature vectors from the previous steps. Specifically, the fused feature vector  $F$ , obtained through the co-attention mechanism, is integrated into the recommendation process by concatenating it with the user-item interaction embedding  $E_M^a$ . This concatenation creates a comprehensive feature vector that captures both the latent user-item interactions and the complementary information between review texts and images. The concatenated feature vector is then fed into the MLP for final preference prediction as follows:



$$\begin{aligned}
 V &= [E_M \oplus F], \\
 V_1 &= \sigma(W_1^V V + b_1^V), \\
 &\dots \\
 V_M &= \sigma(W_M^V V_{M-1} + b_M^V),
 \end{aligned} \tag{12}$$

where  $W_M^V$  and  $b_M^V$  indicate weight matrix and bias, and the output  $V_M$  represents the final vector for the prediction. Finally, we predict the user's preference for a specific item as:

$$\hat{y}_{u,i} = f(W_V V_M), \tag{13}$$

where  $W_V$  indicates the weight matrix for the prediction layer and  $\hat{y}_{u,i}$  indicates predicted user preference, which represents the rating used in this study. Since the predicted rating is derived from regression, we use Mean Squared Error (MSE) as the loss function to conduct an effective learning process as follows:

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^N (y_{u,i} - \hat{y}_{u,i})^2, \tag{14}$$

where  $y_{u,i}$  indicates the actual rating and  $N$  indicates the number of training data. Meanwhile, we use the adaptive moment estimation (Adam) as the optimization method.

#### 4. Experiments

In this study, we conducted extensive experiments using two publicly available datasets from the e-commerce platform Amazon. To verify the performance of the proposed model, we will answer the following research questions (RQs).

- RQ 1: Does the proposed CAMRec model provide better recommendation performance compared to other baseline models?
- RQ 2: How do the fused features of texts and images impact recommendation performance?
- RQ 3: Which fusion method is the most effective in fusing texts and images?

##### 4.1. Datasets

This study uses two datasets from the Amazon as the experimental datasets: Cell Phones and Accessories, and Electronics ([https://cseweb.ucsd.edu/~jmcauley/datasets/amazon\\_v2/](https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/), accessed on 10 May 2024). Amazon datasets are widely used in various recommender system studies as they contain vast amounts of purchase records, as well as review texts and images. To conduct experiments effectively, this study filters and uses user reviews simultaneously containing text and image data. Table 1 summarizes the detailed statistics of the datasets used in this study. We randomly divided 70% for training, 10% for validation, and 20% for test sets [1].

**Table 1.** Statistics of the experimental datasets.

Feature	Cell Phones and Accessories	Electronics
User	148,405	262,488
Item	60,665	102,241
Review and Rating	179,249	344,013
Sparsity (%)	99.998%	99.999%

##### 4.2. Evaluation Metric

To evaluate the performance of our model, we use Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which are widely used metrics in recommender systems. The calculation of MAE is shown in Equation (14), where the absolute differences between predicted and actual ratings are summed and then divided by the total number of evaluated

subjects. Each error contributes equally to the final measure, regardless of its magnitude. Meanwhile, *RMSE* is calculated as in Equation (15), which is the squared error between the predicted and actual ratings that is divided by the total number of subjects under evaluation. Compared to *MAE*, *RMSE* has a relatively larger weight to the larger error value.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_{u,i} - \hat{y}_{u,i}|, \quad (15)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_{u,i} - \hat{y}_{u,i})^2}. \quad (16)$$

#### 4.3. Baseline Models

To validate the trustworthiness of the proposed CAMRec model, we selected the following baseline models for comparison. These have been widely used in recommender system studies. Here are the explanations for each baseline model.

- PMF [28]: This model predicts ratings by modeling latent factors of the user and the item using the evaluation matrix as input based on the Gaussian distribution. This model is effective on sparse, imbalanced rating data.
- NeuMF [29]: This model combines the Generalized Matrix Factorization (GMF) and MLP to measure the nonlinear relationship between the user and item latent factors.
- DeepCoNN [6]: This model uses two CNN processors for each user's review and item's review to extract features, which are combined with Factorization Machine (FM) to predict ratings.
- RSBM [7]: This model extracts the semantic features of review texts using CNN and self-attention mechanism, which predicts the rating based on the importance of each extracted feature.
- VBPR [18]: This model extracts visual representations from item images and incorporates them into the MF model. It can solve the cold-start problems and provide accurate recommendations using Bayesian Personalized Ranking (BPR).
- UCAM [30]: This model predicts ratings by integrating context information into the NeuMF model. In this study, we use RoBERTa and VGG-16 models to extract feature representations from the review texts and images, and then integrate them as context information into the model.

#### 4.4. Experimental Settings

To effectively compare the proposed CAMRec model with the baseline models, we trained our model using the training dataset, determined the optimal hyperparameter values using the validation dataset, and measured the recommendation performance using the test dataset. This study set early stopping if the validation loss did not decrease for five iterations to prevent overfitting, and the results from these experiments are reported as the averaged value of five times. We conducted experiments using 128.0 GB RAM and NVIDIA V100 GPU.

The batch size, learning rate, embedding sizes, and the number of multi-heads were adjusted to find the optimal hyperparameters of the proposed CAMRec model. Specifically, we found the optimal value for each hyperparameter from the following range: [64, 128, 256, 512] for batch size, [0.001, 0.005, 0.0001, 0.0005] for learning rate, [32, 64, 128, 256] for embedding size of user and item, and [2, 4, 6, 8, 10, 12] for the number of multi-heads. For the Cell Phones and Accessories dataset, we determined 256, 0.001, 128, and 6 as the optimal hyperparameter for the batch size, learning rate, embedding size, and the number of multi-heads, respectively. For the Electronics dataset, we determined 256, 0.005, 128, and 4 as the optimal hyperparameters. As for the hyperparameters of the baseline models, we empirically determined their optimal parameter settings by referring to the authors' original papers. Meanwhile, we applied the same settings in our experiments if

some hyperparameters were not specified. As for the pre-trained RoBERTa and VGG-16 models, we did not perform fine-tuning for the experimental datasets. The textual and visual features used in this study were extracted directly from these pre-trained models, leveraging their ability to capture general language and visual patterns.

## 5. Experimental Results

### 5.1. Performance Comparison to Baseline Models (RQ 1)

We compare the recommendation performance of the proposed CAMRec model with the baseline models. The results from Table 2 indicate that the CAMRec outperforms the baseline models in all datasets. Next, we provide insights into three different aspects.

**Table 2.** Performance comparison with baseline models.

Model	Cell Phones and Accessories		Electronics	
	MAE	RMSE	MAE	RMSE
PMF	1.839	2.093	1.795	2.183
NeuMF	1.543	1.675	1.299	1.529
DeepCoNN	0.731	0.971	0.657	0.896
RSBM	0.566	0.830	0.525	0.783
VBPR	1.490	1.637	1.258	1.475
UCAM	0.516	0.804	0.463	0.790
<b>CAMRec</b>	<b>0.460</b>	<b>0.725</b>	<b>0.417</b>	<b>0.701</b>

First, the models using rating as the sole information (e.g., PMF and NeuMF) show the lowest recommendation performance among all baseline models. Such results indicate that a recommendation method using only rating information has limitations in capturing the specific purchase motives of users. However, since NeuMF leverages the nonlinearity of deep neural networks, it can model complex interactions between users and items that linear models like PMF cannot.

Second, models using online reviews (e.g., DeepCoNN, RSBM, VBPR, and UCAM) show better performance than models using only the rating information. The results confirm the improvement in recommendation performance as online reviews contain more information on users' preferences for a specific item. Meanwhile, among the models using review text information, RSBM outperforms DeepCoNN. RSBM uses the target user's review text of an item, while DeepCoNN uses all review texts to learn the user's latent expressions that overlook the evaluation of the item varies with the user. Therefore, the results suggest that learning the target user's review text is more effective for personalized recommendations.

Third, the multimodal model such as UCAM shows better recommendation performance than the models using a single source of information. Multimodal models can minimize the information loss by utilizing both texts and images as the sources of information. Therefore, such results suggest simultaneously using textual and visual information can minimize information loss and ensure accurate recommendations.

Finally, our proposed CAMRec model show the best recommendation performance out of all baseline models for the following reasons: (1) Unlike PMF and NeuMF models using rating information to capture the interaction between users and items, the CAMRec model enhances the recommendation performance using various preference representations inherently found in online reviews. (2) Unlike DeepCoNN, RSBM, and VBPR models that incorporate a single modality, the proposed CAMRec model uses both textual and visual representations to learn. Therefore, the CAMRec model can provide recommendations through a more accurate reflection of diverse user preferences that a single modality may not capture. (3) Unlike the UCAM model which simply concatenates images and texts, the CAMRec model applies a co-attention mechanism to model the complementarity between the two modalities. Therefore, its recommendation performance is expected to improve as the model introduces the complementary and substitutive effects of texts and images.

### 5.2. Effect of Components of CAMRec (RQ 2)

The CAMRec model uses fused textual and visual features to ensure more accurate recommendations. In this section, ablation studies are conducted to verify whether the fused features enhance the recommendation performance. CAM-R is the model that only uses rating information. CAM-RT is the model that uses both rating and textual information from review texts. CAM-RI is the model that uses both rating and visual information from review images.

As suggested in Table 3, models using online reviews outperform the model using the rating in all datasets. This indicates that review-based representations can effectively improve recommendation performance as online reviews contain more specific information about users' preferences and item features. Therefore, using online reviews is crucial for modeling the preference features for users and items. Moreover, the composition of CAM-RI shows lower performance than the one of CAM-RT. The result suggests that written language is a clearer way of expressing users' specific opinions or evaluations than images, thereby acquiring more accurate information about the item. Finally, the CAMRec model shows better recommendation performance than all other variants. It leads to enhanced performance in the case of fusing rating and visual and textual information, which implies the complementarity of texts and images and enables the learning of better comprehensive representations for users and items.

**Table 3.** Effect of components of the CAMRec model.

Model	Cell Phones and Accessories		Electronics	
	MAE	RMSE	MAE	RMSE
CAM-R	1.846	2.220	1.252	1.487
CAM-RT	0.504	0.764	0.479	0.735
CAM-RI	1.334	1.512	1.174	1.389
CAMRec	0.460	0.725	0.417	0.701

### 5.3. Effect of Fusion Strategy (RQ 3)

In this study, we apply the co-attention mechanism to fuse textual and visual features based on complementarity modeling. Therefore, we conducted an experiment to verify which fusion method can be the representative method for fusing the features. Specifically, we determined the addition, average, multiplication, and concatenation operations as the comparison, which are widely used in previous studies. Here, the multiplication operation is performed element-wise. The experimental results are shown in Table 4.

**Table 4.** Effect of fusion strategy on the CAMRec model.

Method	Cell Phones and Accessories		Electronics	
	MAE	RMSE	MAE	RMSE
Addition	0.515	0.763	0.470	0.707
Average	0.553	0.797	0.473	0.711
Multiplication	0.530	0.799	0.486	0.728
Concatenation	0.509	0.755	0.469	0.707
Co-attention	0.460	0.725	0.417	0.701

The experimental results confirm that using the co-attention mechanism provides the best performance, as it leverages the attention mechanism to simultaneously learn the interactions between different modalities and fuse such information by weighting the degree of influence of each modality to the others. However, other operations fuse the modalities, overlooking the complementary and substitutive effects between them, thereby suppressing the performance.

## 6. Conclusions and Future Work

To solve the data sparsity issues in the recommender systems, previous studies have extracted various features from online reviews to integrate them into models. Since online reviews contain rich and specific user preference information, they enhance recommendation performance with clear motivations. Although online reviews contain multimodal information, including texts and images, many studies focus solely on textual information, which overlooks the informational value of images. Regarding the notion of each modality providing different information, previous studies have been limited in considering the complementary or substitutive effect of different modalities. To target this research gap, this study proposes a novel recommender system with a co-attention mechanism to incorporate the complementarity between texts and images. The experimental results were evaluated using two datasets from Amazon, and the proposed CAMRec model outperformed baseline models. This implies that the complementary fusion of texts and images significantly contributes to enhancing recommendation power. Moreover, the CAMRec model is proposed to have scalability, and can efficiently process large-scale datasets with superior performance. This ensures the model is well-suited for real-world applications, specifically in environments where data volumes continuously increase. Therefore, this study provides a new perspective in the field of recommender systems and demonstrates that our proposed model enhances recommendation performance while being adaptable to scalable and large-scale implementations.

The theoretical implications of this study are as follows. First, this study attempted to fuse textual and visual features, which has not yet been performed in previous studies. Therefore, this study provides a new direction and extends the scope of recommender system-related studies. Second, this study provides significant insights in determining how visual information from images plays a vital role in reflecting users' preferences in multimodal studies and in determining an ideal image processing method for the utmost recommendation performance. Third, this study utilizes a co-attention mechanism for modeling the complex interactions between text and images. Therefore, this study lays the theoretical foundation for the exploitation of multimodal data, providing contributions in the fields of multimodal learning as well as deep learning, and opening new possibilities for how to effectively fuse various data sources.

The practical implications of this study are as follows. First, since the proposed CAMRec model can provide more sophisticated and personalized recommendations, users' experience will significantly improve. This may contribute to increasing customer satisfaction and user retention in e-commerce platforms. Second, the proposed CAMRec model is designed with domain-agnostic characteristics, making it highly scalable and adaptable across various domains. Although this study uses datasets from Amazon, which is common in recommender system studies, the model's architecture is flexible enough to be applied to other domains such as fashion, entertainment, or social media platforms. Since each domain involves multimodal data (e.g., texts and images), the CAMRec model can process data through its co-attention mechanism. Moreover, the scalability of the model supports large-scale datasets and growing operational demands, making it well-suited for providing efficient recommendations in contexts beyond e-commerce. Finally, the feature composition of our proposed CAMRec model is model-agnostic, which allows for flexible implementation. Therefore, businesses can easily apply it to their existing hardware or software environments without major overhauls. Such flexibility enables companies to overcome various technical constraints while optimizing the model for their specific operational needs. Moreover, companies can enhance customer experience and business performance by incorporating the insights gained from this study.

Although the effectiveness of this study has been demonstrated in various perspectives, some limitations still exist. First, this study uses RoBERTa and VGG16 models to extract textual and visual information from online reviews. Besides these two models, other existing models are also used for feature extraction in studies. Therefore, we can conduct comparisons in future works by using different models to verify whether the extraction

model impacts recommendation performance. Second, the experimental datasets used in this study are collected from Amazon, which limits the evaluation to the e-commerce domain. Since the proposed CAMRec model is designed with domain-agnostic characteristics, the effectiveness and generalizability in other domains need to be verified. Besides e-commerce, other domains also have different types of multimodal data and user behaviors. Therefore, we can explore the applicability of the CAMRec model across diverse domains in future works. Third, this study approaches the recommender system using multimodal data, including review texts and images. Besides these, different pieces of auxiliary information, such as titles and descriptions of items, have also widely been used in previous studies and impact recommendation performance. Therefore, we can use extra auxiliary information in future works to verify if it improves recommendation performance. Fourth, this study applies co-attention to achieve the purpose of introducing the complementarity between texts and images in the recommendations. As for the fusion strategy, some studies have proposed other advanced techniques, such as cross-attention mechanisms, gated multimodal units, multimodal factorized bilinear pooling, etc. Therefore, we can compare advanced fusion techniques in future works to verify whether the applied co-attention mechanism is effective in recommendation performance. Finally, although this study demonstrates the potential of multimodal data in improving recommendation performance, the cold-start issues commonly present in recommender systems have not been specifically addressed. Since this study leverages rich multimodal data, we believe it can mitigate cold-start issues with detailed descriptive content for items. Therefore, we can analyze whether the proposed CAMRec model better addresses cold-start scenarios in future works.

**Author Contributions:** Conceptualization, E.J., A.K. and J.K.; Methodology, X.L. and S.P.; Software, E.J. and Q.L.; Data curation, E.J. and S.P.; Writing—original draft, E.J. and A.K.; Writing—review and editing, X.L., Q.L. and J.K.; Supervision, J.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the BK21 FOUR Program (5199990913932) funded by the Ministry of Education (MOE, Republic of Korea) and the National Research Foundation of Korea (NRF).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data are available on [https://cseweb.ucsd.edu/~jmcauley/datasets/amazon\\_v2/](https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/) (accessed on 10 May 2024).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Jang, D.; Li, Q.; Lee, C.; Kim, J. Attention-based multi attribute matrix factorization for enhanced recommendation performance. *Inf. Syst.* **2024**, *121*, 102334. [\[CrossRef\]](#)
2. Zhu, Z.; Yan, M.; Deng, X.; Gao, M. Rating prediction of recommended item based on review deep learning and rating probability matrix factorization. *Electron. Commer. Res. Appl.* **2022**, *54*, 101160. [\[CrossRef\]](#)
3. Park, J.; Li, X.; Li, Q.; Kim, J. Impact on recommendation performance of online review helpfulness and consistency. *Data Technol. Appl.* **2023**, *57*, 199–221. [\[CrossRef\]](#)
4. Li, Q.; Li, X.; Lee, B.; Kim, J. A hybrid CNN-based review helpfulness filtering model for improving e-commerce recommendation Service. *Appl. Sci.* **2021**, *11*, 8613. [\[CrossRef\]](#)
5. Liu, H.; Wang, Y.; Peng, Q.; Wu, F.; Gan, L.; Pan, L.; Jiao, P. Hybrid neural recommendation with joint deep representation learning of ratings and reviews. *Neurocomputing* **2020**, *374*, 77–85. [\[CrossRef\]](#)
6. Zheng, L.; Noroozi, V.; Yu, P.S. Joint deep modeling of users and items using reviews for recommendation. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Cambridge, UK, 6–10 February 2017; pp. 425–434.
7. Cao, R.; Zhang, X.; Wang, H. A review semantics based model for rating prediction. *IEEE Access* **2019**, *8*, 4714–4723. [\[CrossRef\]](#)
8. Liu, Y.-H.; Chen, Y.-L.; Chang, P.-Y. A deep multi-embedding model for mobile application recommendation. *Decis. Support Syst.* **2023**, *173*, 114011. [\[CrossRef\]](#)



9. Xu, C.; Guan, Z.; Zhao, W.; Wu, Q.; Yan, M.; Chen, L.; Miao, Q. Recommendation by users' multimodal preferences for smart city applications. *IEEE Trans. Ind. Inform.* **2020**, *17*, 4197–4205. [\[CrossRef\]](#)
10. Ren, G.; Diao, L.; Guo, F.; Hong, T. A co-attention based multi-modal fusion network for review helpfulness prediction. *Inf. Process. Manag.* **2024**, *61*, 103573. [\[CrossRef\]](#)
11. Xiao, S.; Chen, G.; Zhang, C.; Li, X. Complementary or substitutive? A novel deep learning method to leverage text-image interactions for multimodal review helpfulness prediction. *Expert Syst. Appl.* **2022**, *208*, 118138. [\[CrossRef\]](#)
12. Liu, M.; Liu, L.; Cao, J.; Du, Q. Co-attention network with label embedding for text classification. *Neurocomputing* **2022**, *471*, 61–69. [\[CrossRef\]](#)
13. Yang, S.; Li, Q.; Jang, D.; Kim, J. Deep learning mechanism and big data in hospitality and tourism: Developing personalized restaurant recommendation model to customer decision-making. *Int. J. Hosp. Manag.* **2024**, *121*, 103803. [\[CrossRef\]](#)
14. Takács, G.; Pilászy, I.; Németh, B.; Tikk, D. Scalable collaborative filtering approaches for large recommender systems. *J. Mach. Learn. Res.* **2009**, *10*, 623–656.
15. Ma, Y.; Chen, G.; Wei, Q. Finding users preferences from large-scale online reviews for personalized recommendation. *Electron. Commer. Res.* **2017**, *17*, 3–29. [\[CrossRef\]](#)
16. Chen, C.; Zhang, M.; Liu, Y.; Ma, S. Neural attentional rating regression with review-level explanations. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 1583–1592.
17. Liu, P.; Zhang, L.; Gulla, J.A. Dynamic attention-based explainable recommendation with textual and visual fusion. *Inf. Process. Manag.* **2020**, *57*, 102099. [\[CrossRef\]](#)
18. He, R.; McAuley, J. VBPR: Visual bayesian personalized ranking from implicit feedback. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
19. Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; Zha, H. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 765–774.
20. Liu, F.; Chen, H.; Cheng, Z.; Liu, A.; Nie, L.; Kankanhalli, M. Disentangled multimodal representation learning for recommendation. *IEEE Trans. Multimed.* **2022**, *25*, 7149–7159. [\[CrossRef\]](#)
21. Huang, F.; Zhang, X.; Zhao, Z.; Xu, J.; Li, Z. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowl.-Based Syst.* **2019**, *167*, 26–37. [\[CrossRef\]](#)
22. Zhang, W.; Yu, J.; Zhao, W.; Ran, C. DMRFNet: Deep multimodal reasoning and fusion for visual question answering and explanation generation. *Inf. Fusion* **2021**, *72*, 70–79. [\[CrossRef\]](#)
23. Ren, G.; Diao, L.; Kim, J. DMFN: A disentangled multi-level fusion network for review helpfulness prediction. *Expert Syst. Appl.* **2023**, *228*, 120344. [\[CrossRef\]](#)
24. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 289–297.
25. Laenen, K.; Moens, M.-F. A comparative study of outfit recommendation methods with a focus on attention-based fusion. *Inf. Process. Manag.* **2020**, *57*, 102316. [\[CrossRef\]](#)
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
28. Mnih, A.; Salakhutdinov, R.R. Probabilistic matrix factorization. *Adv. Neural Inf. Process. Syst.* **2007**, *20*, 1257–1264.
29. He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; Chua, T.-S. Neural collaborative filtering. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 173–182.
30. Unger, M.; Tuzhilin, A.; Livne, A. Context-aware recommendations based on deep learning frameworks. *ACM Trans. Manag. Inf. Syst.* **2020**, *11*, 1–15. [\[CrossRef\]](#)

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.