sensors

MDPI

Communication

CMBF: Cross-Modal-Based Fusion Recommendation Algorithm

Xi Chen, Yangsiyi Lu, Yuehai Wang * and Jianyi Yang

College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310063, China; chen__xi@zju.edu.cn (X.C.); luyangsiyi@zju.edu.cn (Y.L.); yangjy@zju.edu.cn (J.Y.)
* Correspondence: wangyuehai@zju.edu.cn

Abstract: A recommendation system is often used to recommend items that may be of interest to users. One of the main challenges is that the scarcity of actual interaction data between users and items restricts the performance of recommendation systems. To solve this problem, multi-modal technologies have been used for expanding available information. However, the existing multi-modal recommendation algorithms all extract the feature of single modality and simply splice the features of different modalities to predict the recommendation results. This fusion method can not completely mine the relevance of multi-modal features and lose the relationship between different modalities, which affects the prediction results. In this paper, we propose a Cross-Modal-Based Fusion Recommendation Algorithm (CMBF) that can capture both the single-modal features and the cross-modal features. Our algorithm uses a novel cross-modal fusion method to fuse the multi-modal features completely and learn the cross information between different modalities. We evaluate our algorithm on two datasets, MovieLens and Amazon. Experiments show that our method has achieved the best performance compared to other recommendation algorithms. We also design ablation study to prove that our cross-modal fusion method improves the prediction results.

Keywords: recommendation systems; multi-modal algorithm; cross-modal fusion; attention mechanism

check for updates

Citation: Chen, X.; Lu, Y.; Wang, Y.; Yang, J. CMBF: Cross-Modal-Based Fusion Recommendation Algorithm. *Sensors* **2021**, *21*, 5275. https://doi.org/10.3390/s21165275

Academic Editors: Yinong Chen and Steven Guan

Received: 8 June 2021
Accepted: 29 July 2021
Published: 4 August 2021

1. Introduction

The rapid development of the information age is a double-edged sword which not only brings convenience to people's lives but also brings us troubles such as data flooding problem. Imagine that when we just want to choose a movie for entertainment without a clear goal, we will be likely to become lost in countless movies. We need an "expert" that can automatically analyze our historical interests and find the specific movie meeting our individual needs from a large library of movies. This "expert" is the recommendation system [1].

Recommendation systems typically record the preference of a user according to the user's historical operation, and then recommend according items that the user may also like. In the recommendation system, the number of users and recommended items is large, but the number of actual interactions between the users and items is usually rare, although the interaction data are important for recommendation. Many experts have made many efforts to solve the sparsity problem of the interaction data. Some studies [2,3] indicate that a content-based recommendation algorithm can calculate the similarity of the users or items through characteristics to ease the data sparsity problem. For example, Gunawardana and Meek [4] use Boltzmann Machines to model the relationship between the users and items. They bound the parameters of the model with content features to solve the data sparsity problem. Later, more content-based algorithms using deep neural networks are proposed to learn the content features more effectively [5–8]. However, these algorithms cannot work well when the original characteristics of the users and items are not enough.

The importance of image features in recommendation systems: Recently, multi-modal technologies have been used in recommendation systems [9–11] to extend the properties of the users and items by using complementary information such as images. The image feature of the item provides some information that is not available in the text, which has an impact on the user's preference. For example, one would not buy a t-shirt from Amazon without seeing an image of the item [12]. In addition, in movie recommendation, the user can obtain an more intuitive understanding of the movie through the images, which can help to determine whether he/she is interested in the movie. Therefore, incorporating image features into the recommendation factors can have a positive impact on the accuracy of the recommended results.

Most of the existing multi-modal recommendation algorithms use the same feature learning method and do not additionally learn the relationship between different modal features, which could lose some fine-grained information and affect the prediction results.

In this paper, we propose a Cross-Modal-Based Fusion Recommendation Algorithm (CMBF) to alleviate the data sparsity problem mentioned above. Our algorithm can capture both the single-modal feature and the cross information between two modal features. The CMBF contains four modules: the preprocessing module, single-modal learning module, cross-modal fusion module and output module. Firstly, the image and text features are extracted and embedded in the preprocessing module. Next, we use the single-modal learning module to learn the feature of the single modality. The single-modal features are then fed into the cross-modal fusion module to learn the cross-modal features between two modalities. Finally, we fuse both the single-modal features and the cross-modal features to obtain the high-level feature and predict the recommendation results. An attention mechanism is used in our algorithm to extract more meaningful feature. We also adopt residual connection to combine different levels of feature information.

To summarize, in this paper we make the following contributions:

- We find the problem that insufficient fusion of multi-modal features will lead to the loss of some fine-grained information and affect the prediction results;
- We propose a novel cross-modal fusion method to fuse the multi-modal features completely and learn the cross information between different modalities;
- We evaluate our algorithm on two datasets, MovieLens and Amazon. Experiments show that our method has achieved the best performance compared to other recommendation algorithms.

The rest of the paper is organized as follows. Section 2 presents the problem definition of the recommendation system and presents the related works. Section 3 presents the model architecture of the CMBF. Section 4 presents the experiments and results on the MovieLens dataset and Amazon dataset. Section 5 concludes this paper and points out the future work.

2. Related Work

People can gain information from images, sounds, texts and so on. In other words, our world is a multi-modal world. When a research question or dataset contains multiple modalities, it can be processed by multi-modal technologies. Multi-modal technologies can be applied in various fields. For example, one of the earliest applications of multimodal research is audiovisual speech recognition (AVSR) [13], which uses visual information to improve the accuracy of speech recognition. Multi-modal technologies have also played an important role in emotion recognition [14], image description [15], VQA [16], traffic event detection [17] and other fields.

Using multi-modal technologies in the recommendation system can alleviate the data sparsity problem and optimize the performance of the recommendation system. In 2007, Microsoft's Yang et al. [18] first proposed the concept of multi-modal recommendation. They use the three-modal information of text, image and audio as input to calculate the similarity of between each pair of modals separately, and then use Attention Fusion Function for fusion. Although the feature extraction of the three modalities is very rough at that time, this article still plays a milestone role in the research of multi-modal recommendation algorithms, and the subsequent work can be expended on this basis. For instance, Oramas et al. [19] use the artist's text to describe information and audio track information to solve the cold start problem in music recommendation. They aggregate all the songs of the same artist to learn the characteristics of the artist, then learn the music feature information for the sound track, and stitch them as the input of a multi-modal fusion network. Finally they get the fusion representation feature of the entire music; Cai et al. [9] of Youku proposed a multi-view active learning framework for video recommendation, which extracts missing text information from visual information to obtain more training videos; Ge et al. [10] believe that the use of user behavior images to enhance behavior representation is helpful to understand the user's visual preferences and greatly improve the accuracy of prediction. Therefore, they propose to jointly model user preferences with user behavior ID characteristics and behavior images; Wu et al. [11] proposed a multi-view news recommendation based on the attention mechanism. They regard information such as headlines, texts, and subject classifications as multiple forms of news. They encode different forms of information, and then use the attention mechanism to fuse them. These multi-modal recommendation methods usually extract the feature of single modality and simply concatenate the features of different modalities to predict the recommendation results. Obviously, the simple concatenation cannot fuse the multi-modal features completely which may cause the loss of important information. Therefore, we propose a new fusion method called cross-modal fusion to fully fuse the multi-modal features and achieve fine-grained information for prediction. The details of the cross-modal fusion method are presented in Section 3.5.

3. Model Architecture

3.1. Problem Definition

In a classic recommendation system, we define a user set $U = \{u_1, u_2, \ldots, u_N\}$ that represents N users, where u_i represents the i-th user. The feature of each user u can be described by a feature vector $X_u = \{x_1, x_2, \ldots, x_P\}$, where x_j represents the j-th feature and P represents the total feature number. Similarly, we can also define an item set $V = \{v_1, v_2, \ldots, v_M\}$ to represent M items, with the feature vector $X_v = \{x_1, x_2, \ldots, x_Q\}$ representing the feature of each item v. The interaction matrix between the user and the item is defined as $Y = \{y_{uv} | u \in U, v \in V\}$, where y_{uv} represents the interaction of the user u with the item v obtained according to the user's implicit feedback, such as clicking, browsing, purchasing and scoring. The value of y_{uv} is set to 1 when the interaction between the user u and the item v is observed, otherwise the value of y_{uv} is 0. The goal of the recommendation system is to infer the possible interaction $\hat{y}_{uv}$ between the user u and the item v that have not been shown before.

3.2. Overview

Our CMBF framework can be divided into four parts, as shown in Figure 1. Firstly, we extract the image and text feature of users and items and encode them through different embedding methods in the preprocessing module. Then, the dimensionality-reduced Image Embedding Vectors and Text Embedding Vectors are sent to the single-modal learning module, respectively. Next, the interaction and fusion of the two modalities is performed in the cross-modal fusion module to learn the high-level feature representation. Finally, in the output module, the output of the cross-modal fusion module will be further used to predict the recommendation result. The specific implementation details will be expanded in subsequent sections.

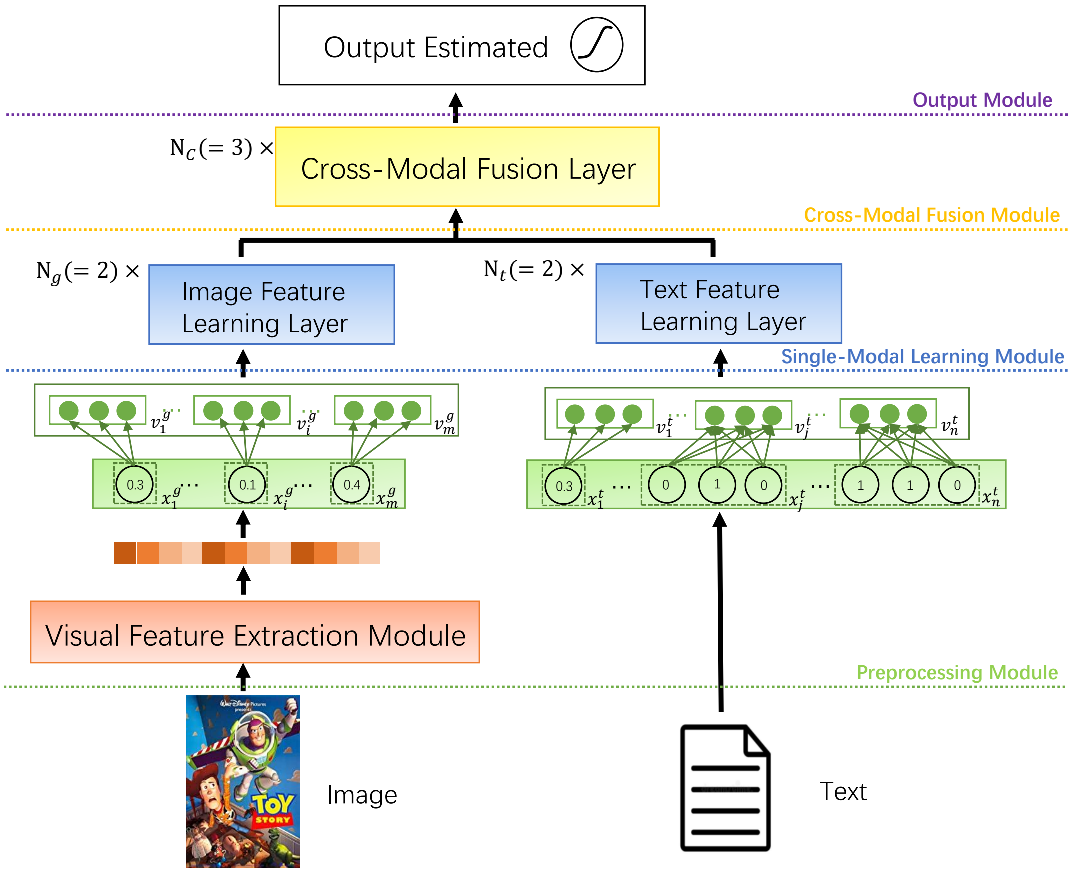

Figure 1. Overview of the proposed framework based on CMBF. The details of the Image/Text Feature Learning Layer and the Cross-modal Fusion Layer are illustrated in Figures 2 and 3, respectively.

3.3. Preprocessing Module

Firstly, we should extract the image and text feature of the users and items, respectively. As shown in Figure 1, the Visual Feature Extraction Module (generally CNN-based) is used to obtain the item's image feature. The image feature vector of the item v can be denoted as

$$x^{g,v} = \{x_1^{g,v}, x_2^{g,v}, \ldots, x_m^{g,v}\}, \tag{1}$$

where g represents 'image', and m is the length of the image feature vector $x^{g,v}$.

Since the text features include the numerical category and the classification category, we use different methods to obtain the text feature. As shown in Figure 1, we use the specific value directly as the corresponding value of the numerical text feature and use the binary representation of "0/1" to represent the categorized text feature. For example, if the price of an item is "10.5 \$", the text feature of the price is "10.5"; if the brand of the item is "Fossil", the value of the "*Fossil*" is "1" in the text feature of the brand. Then we can obtain the text feature vector of the item v and the user u, respectively as

$$x^{t,v} = \{x_1^{t,v}, x_2^{t,v}, \ldots, x_{v_n}^{t,v}\}, \quad x^{t,u} = \{x_1^{t,u}, x_2^{t,u}, \ldots, x_{u_n}^{t,u}\}, \tag{2}$$

where t represents 'text', v_n and u_n represent the length of the text feature vector $x^{t,v}$ and $x^{t,u}$, and $v_n + u_n = n$, n is the total length of the text feature vector. We combine all the image and text feature, respectively, to obtain the final representation of the two modal feature vectors as follows:

$$x^g = x^{g,v} = \{x_1^g, x_2^g, \ldots, x_m^g\}, \quad x^t = \mathrm{concat}\{x^{t,v}, x^{t,u}\} = \{x_1^t, x_2^t, \ldots, x_n^t\}. \tag{3}$$

To alleviate the high dimensionality and sparseness of the feature vectors, we use the following methods proposed in [7] to perform embedding dimensionality reduction on the numerical feature, single-value classification feature, and multi-value classification feature:

$$v_i = \begin{cases} w_i x_i, & \text{for numerical features,} \\ W_i x_i, & \text{for single} - \text{value classification features,} \\ \frac{1}{Q} W_i x_i, & \text{for multi} - \text{value classification features,} \end{cases} \tag{4}$$

where x_i is the element in the image or text feature vector, v_i represents the reduced-dimensional embedding vector, w_i represents the Embedding Mapping Vector in the case of the numerical feature, W_i represents the Embedding Mapping Matrix in the case of the classification feature, and Q represents the number of all potential values if x_i is a multi-value feature vector.

After the image feature vectors and the text feature vectors are both encoded as Equation (4), the Image Embedding Vectors and the Text Embedding Vectors will be obtained and prepared for the next module.

3.4. Single-Modal Learning Module

The multi-head self attention mechanism [20] can help to learn subtle feature information in the single-modal feature, and each "head" can be regarded as a subspace. Taking the text feature as an example, in the h-th feature subspace, the similarity between the Text Embedding Vector v_i and the Text Embedding Vector v_j can be calculated as

$$\theta^{(h)}(v_i, v_j) = \text{Similarity}(W^{(h)}_{Query} v_i, W^{(h)}_{Key} v_j), \tag{5}$$

where $W^{(h)}_{Query}$ and $W^{(h)}_{Key} \in R^{d' \times d}$ are the transformation matrix that maps the embedding vectors from the original space $R^{d'}$ to the space R^d. Then the weighted representation of v_i in the h-th feature subspace is

$$\tilde{v}^{(h)}_i = \sum_{k=1}^{n} \alpha^{(h)}_{t,i,k} (W^{(h)}_{Value} v_k) = \sum_{k=1}^{n} \frac{exp(\theta^{(h)}_t (v_i, v_j))}{\sum_{k=1}^{m} exp(\theta^{(h)}_t (v_i, v_k))} (W^{(h)}_{Value} v_k). \tag{6}$$

In Equation (6), $\alpha^{(h)}_{t,i,k}$ is the normalized weight of $\theta^{(h)}_t(v_i, v_j)$, $W^{(h)}_{Value} \in R^{d' \times d}$ is the transformation matrix, and m is the total number of the text feature number. By splicing all the subspaces and using residual connection, we can obtain the new representation of the text embedding vector:

$$v^{Res}_i = \text{ReLU}(\tilde{v}_i + W_{Res} v_i), \tag{7}$$

where v_i is the concatenation of $\{\tilde{v}^{(1)}_i, \tilde{v}^{(2)}_i, \ldots, \tilde{v}^{(H)}_i\}$, H is the total number of the subspaces, and $W_{Res} \in R^{Hd' \times d}$ is the mapping matrix.

As shown in Figure 2, the operation through Equations (5)–(7) is called a feature learning layer which can get feature representation of the text or image. We stack multiple feature learning layers to obtain a single-modal feature learning module, so as to obtain the representation of the single-modal feature.

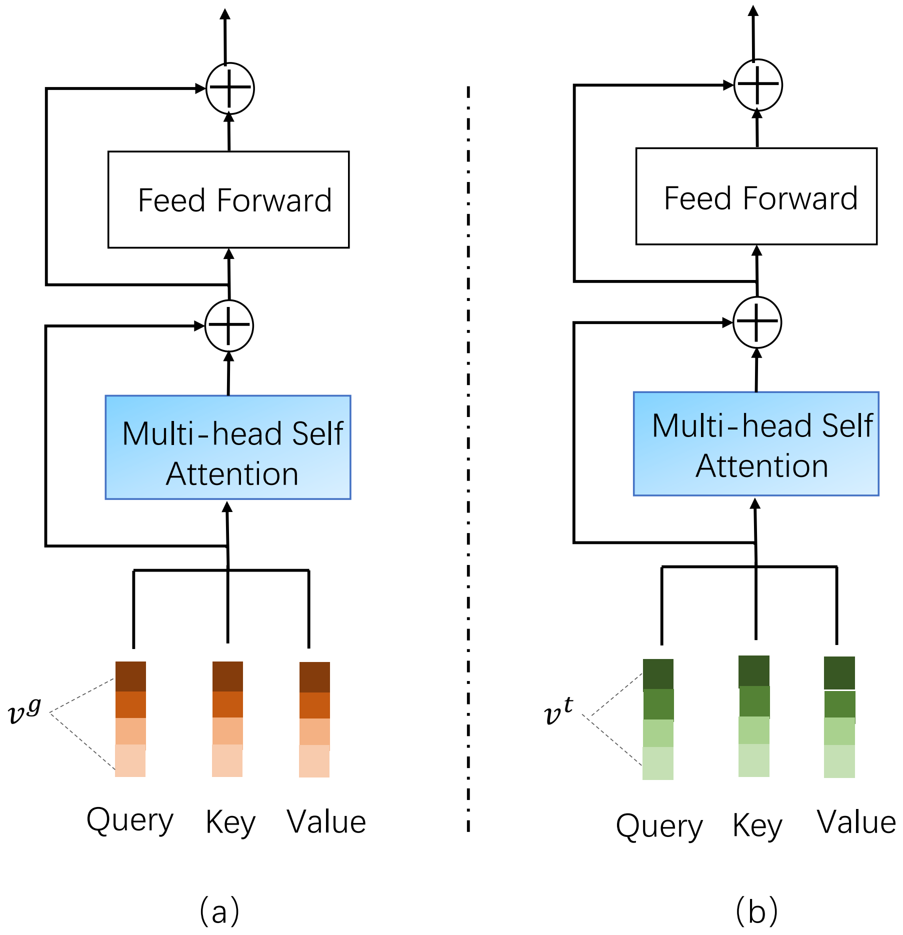

Figure 2. Illustration of the Feature Learning Layer. (**a**) Represents the Image Feature Learning Layer and (**b**) represents the Text Feature Learning Layer.

3.5. Cross-Modal Fusion Module

In addition to the single-modal feature, the relationship between the text modal feature and the image modal feature of items can also be obtained by data mining and used as supplementary information to further alleviate the data sparse problem of the recommendation algorithm. This is also one of the problems that need to be solved in multi-modal learning technology [21].

Compared with existing multi-modal recommendation algorithms that simply splice single-modal recommendation results, we try to exchange the feature information of the two modalities before fusing these two modal features, and then continue to learn high-level representations of feature. As this feature information exchange operation is similar to a mutual reference of feature across two modalities, we call it "cross-modal fusion".

We denote the output of the text feature and the image feature after their respective single-modal learning modules as: e^t and e^g. The cross-modal fusion layer is composed of two cross attention layers and two self-attention [20] layers as shown in Figure 3. We use the image feature e^g to calculate the cross-modal weight $\alpha_k^{g\to t}$ of the text feature e^t, and obtain the text cross representation e^t_{Cross} using $\alpha_k^{g\to t}$ and e^t:

$$\alpha_k^{g\to t} = \text{softmax}\left(\frac{e_k^g {e_k^t}^T}{\sqrt{d_{e_k^t}}}\right), \tag{8}$$

$$e^t_{Cross} = \sum_{k=1}^{n} \alpha_k^{g\to t} e_k^t. \tag{9}$$

In Equation (8), softmax$(\cdot)$ is a normalization function, $d_{e_k^t}$ is the dimension of e_k^t, and $e_k^g {e_k^t}^T$ represents that the relationship between two modal feature is learned through the inner product. In the same way, we can use the text feature e^t to calculate the cross-modal weight $\alpha_k^{t\to g}$ of the image feature e^g, and obtain the image cross representation e^g_{Cross} using $\alpha_k^{t\to g}$ and e^g:

$$\alpha_k^{t\to g} = \text{softmax}(\frac{e_k^t {e_k^g}^T}{\sqrt{d_{e_k^g}}}), \tag{10}$$

$$e^g_{Cross} = \sum_{k=1}^{n} \alpha_k^{t\to g} e_k^g. \tag{11}$$

Then we send the text cross representation e^t_{Cross} and the image cross representation e^g_{Cross} to the self-attention layer and use the residual connection to maintain the previous feature information. To obtain the higher-level information, we stack multiple cross-modal fusion layers to calculate the relationship between the two modal features, and finally concatenate all the feature vectors for the subsequent prediction.

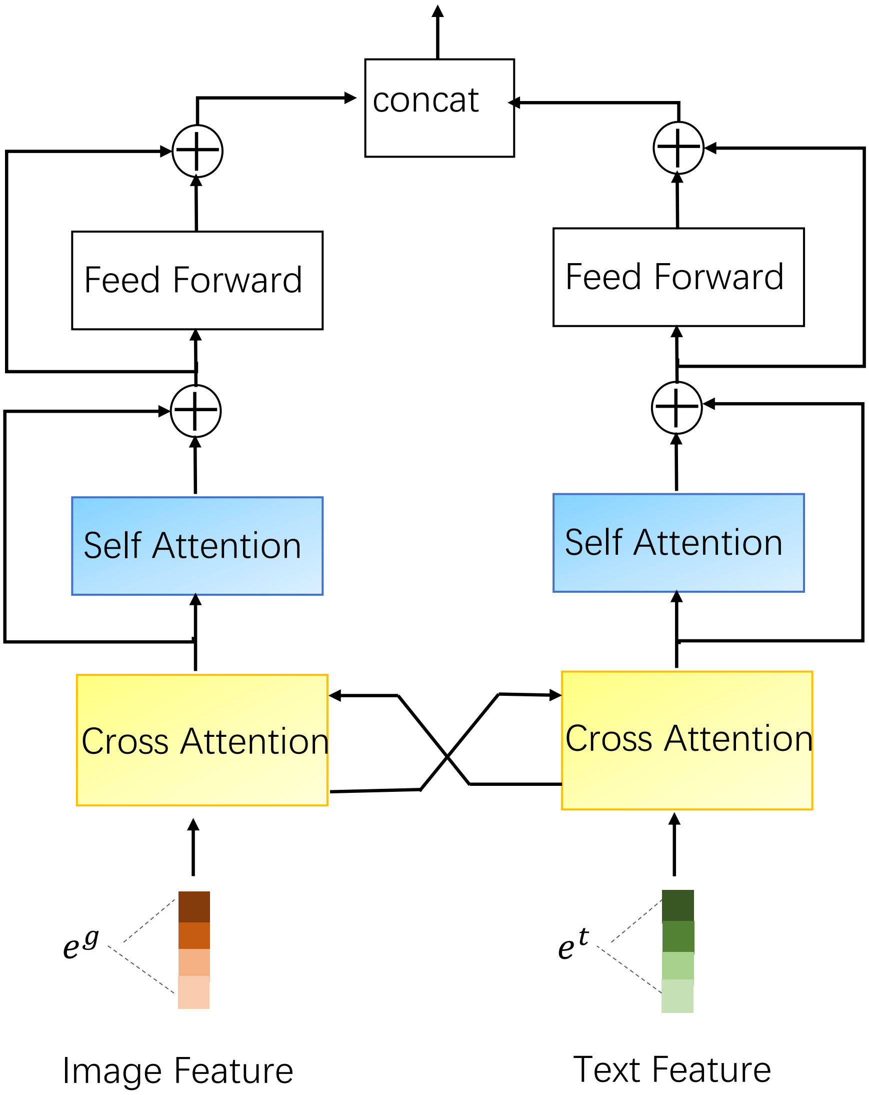

Figure 3. Illustration of the Cross-modal Fusion Layer.

4. Experiments and Results

4.1. Evaluated Datasets

The datasets used in this experiment are the MovieLens dataset [22] and Amazon dataset [23]. The MovieLens dataset is a public movie data set, which contains multiple users' rating information for different movies, as well as user and movie characteristics information. It has multiple data scales according to the number of ratings, and we use the movie dataset MovieLens-1M with a sample size of about 1M in this paper. The Amazon Product dataset is a public electronic product dataset, which contains Amazon product reviews and metadata, as well as user ratings and reviews of products. We use a subset of Clothing & Shoes & Jewelry for experimental testing. The introduction of the two data sets is shown in Table 1.

Table 1. Introduction of the MovieLens dataset and the Amazon dataset.

Dataset	User	Item	Interaction	Density
MovieLens	6040	3685	998,034	4.48%
Amazon	39,387	23,033	278,677	0.031%

Both datasets lack relevant image information, but related download links are provided in the original data. That is, the MovieLens dataset provides download links for movie posters, and the Amazon dataset provides download links for product images. We use crawler technologies to obtain the corresponding image of each movie or product, and then we use the InceptionV3 model [24] pre-trained on ImageNet (ILSVRC2012) as the image feature extraction module to extract image features from these pictures. In this experiment, we use the 2048-dimensional features by the third pooling layer of InceptionV3 as the image feature for subsequent model calculations.

The processed attributes of the MovieLens dataset and Amazon dataset are shown in Tables 2 and 3, respectively. For uncommon attributes which appear less than a certain threshold, we mark them as a single attribute as "unknown". In the MovieLens dataset and the Amazon dataset, the thresholds are 10 and 5, respectively. Specifically, for the attributes of movie year, we treat every ten years from 1919 to 1970 as an attribute, every year after 1970 as an attribute, and the rest of the years as "unknown" attribute. In the two datasets, the user's score ranges from 0 to 5 points, so we should binarize the scoring information of the original dataset as classification labels. Therefore, scores greater than 3 are regarded as positive samples and scores below 3 are considered negative. We eliminate neutral samples with scores equal to 3 points. In addition, we divide the entire dataset into a training set, validation set and test set according to the ratio of 8:1:1 for experiments.

Table 2. The processed attributes of the MovieLens dataset.

	Attributes	Dimension
User	Gender	2
	Occupation	21
	Age	61
	Zip Code	795
Movie	Type	19
	Year	37
	Image	2048
Others	Score Time	1

Table 3. The processed attributes of the Amazon dataset.

	Attributes	Dimension
User	ID	39,387
Product	Price	121
	TOP Sales	19
	Sales Rank	150
	Brand	157
	Type	1193
	Image	2048
Others	Score Time	1

4.2. Experimental Setup

4.2.1. Competing Algorithms

We have divided all the competing algorithms into three categories. The first category is the classic algorithm used before the advent of deep networks:

- **LR** [25] is a commonly used model for recommendation tasks before the deep network is proposed.
- **FM** [26] simulates the importance of the first-order feature and the interaction of the second-order feature.

The second category is the single-modal recommendation algorithm implemented using deep networks:

- **DeepFM** [27] is an end-to-end model using a joint decomposition machine and a multi-layer perceptron. It uses deep neural networks and factorization machines to model the interaction of high-order feature and the interaction of the low-order feature, respectively.
- **Wide&Deep** [28] is a hybrid model composed of a single-layer Wide part and a multi-layer Deep part. The main function of the Wide part is to give the model a strong "memory ability"; the main function of the deep part is to give the model a "generalization ability", so that the model has both the advantages of logistic regression and deep neural network.
- **AutoInt** [7] is an encoder that automatically learns high-order feature combinations of input feature. It maps digital features and classification features to the same low-dimensional space and uses a self-attention mechanism to learn the interaction of the feature in the low-dimensional space.
- **MiFiNN** [29] calculates the mutual information of each sparse feature and the click result as the weight of each sparse feature. It constructs an interactive method combining the outer product and inner product to carry out the feature interaction.
- **ADI** [30] captures the latent interest sequence in the interest extractor layer, and employs auxiliary losses to produce the interest state with deep supervision.

The third category is the recommendation algorithm based on multi-modal fusion:

- **VBPR** [12] integrates visual information into the prediction of people's preferences. Compared with the matrix factorization model that only relies on the hidden vector of the user and the item, VBPR is greatly improved.
- **MLFM** [11] is a multimodal late fusion classification method based on text and image. They use machine learning models to extract text and image features, learn a specific classifier for each modal, and then learn a fusion strategy from the results of each modal classifier.
- **DICM** [10] combines user preferences with user behavior ID characteristics and behavior images. This method merges the user's historically visited item pictures and candidate item pictures using an attention mechanism, and finally fully interacts the

traditional ID feature, candidate item image information, and the user's historical visual preferences to obtain the final prediction result.

4.2.2. Parameter Settings

We use an Adam optimizer [31] with $lr = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and set the batch size as 1024, the dimension of the embedding vector as 16. In the single-modal learning module, the number of the image feature learning layers is denoted as N_g and the number of the text feature learning layers is denoted as N_t. In the cross-modal fusion module, the number of the cross-modal fusion layers is denoted as N_c. Following AutoInt [7], we set $N_g = N_t = 2$. After experiments, we find that the performance of CMBF is best when $N_c = 3$. The details of the parameter ablation experiment are described in Section 4.4.1. In each layer, the number of attention heads is two.

4.2.3. Evaluation Metrics

We use Logloss, AUC, GAUC as the evaluation metrics. The calculation formula of Logloss is as follows:

$$\text{Logloss} = -\frac{1}{T}\sum_{t=1}^{T}(y_t log(\hat{y}_t) + (1 - y_t) log(1 - \hat{y}_t)), \tag{12}$$

where y_t and $\hat{y}_t$ are the real and predicted interaction, respectively, T is the total number of the samples in the training set. AUC (Area Under Curve) is a metric which can evaluate the performance of the classification algorithm. However, AUC does not treat different users differently. Then users who do not click on any advertisements may make the AUC results tend to be lower. Therefore, we additionally use GAUC (Group AUC) metric proposed by Alibaba's Gai et al. [32] to evaluate the model. GAUC can be considered as a weighted average of the AUC of all users:

$$\text{GAUC} = \frac{\sum_{u \in U} w_u \times \text{AUC}_u}{\sum_{u \in U} w_u}, \tag{13}$$

where w_u is the number of operations of the user u, AUC_u is the AUC value of the user u.

4.3. Results and Analysis

The experiment results on the MovieLens dataset are shown in Table 4. Our algorithm CMBF achieves the best performance compared to other recommendation algorithms. The performance of CMBF is much better than the classic algorithms (i.e.; LR and FM) since LR and FM cannot capture non-linear relations of information. Compared with the single-modal models (i.e.; DeepFM, Wide&Deep and AutoInt), CMBF uses multi-modal information to alleviate data sparsity and better capture user preference information for more personalized recommendations. The difference between the CMBF algorithm and other multi-modal algorithms (i.e.; VBPR, MLFM and DCIM) is that CMBF exchanges the feature information of the two modalities before fusing the two modal features. The original representation of users and items include both single-modality and multi-modality, and now we have added cross-modal information to the representation of users and items through the cross-modal module, so that we can use the learned relationship between the two modal features to further enrich the data and alleviate the problem of data sparseness.

Table 4. Comparison of different algorithms on the MovieLens dataset. The results in bold represent the best performances.

Algorithm	AUC	GAUC	Logloss
LR [25]	0.7775	0.5028	0.5441
FM [26]	0.7777	0.6004	0.5176
DeepFM [27]	0.7809	0.6837	0.4351
Wide&Deep [28]	0.7920	0.7239	0.4683
AutoInt [7]	0.8399	0.7712	0.3839
MiFiNN [29]	0.8772	-	0.3382
ADI [30]	0.8417	-	-
VBPR [12]	0.8419	0.7690	0.3700
MLFM [11]	0.8489	0.7789	0.3731
DCIM [10]	0.8655	0.7868	0.3665
CMBF	**0.8836**	**0.8363**	**0.3302**

The experiment results on the Amazon dataset are shown in Table 5. The overall effect of all methods on the Amazon dataset is not as good as that on the MovieLens dataset. The reason is that the sparsity of the Amazon dataset is much higher than that of the MovieLens dataset. Therefore, high sparsity is indeed a key reason that limits the performance of recommendation algorithms, and it is worthwhile to continue to be explored.

Table 5. Comparison of different algorithms on the Amazon dataset. The results in bold represent the best performances.

Algorithm	AUC	GAUC	Logloss
LR [25]	0.4868	0.4793	0.3734
FM [26]	0.5137	0.4978	0.3524
DeepFM [27]	0.5625	0.5255	0.3426
Wide&Deep [28]	0.6264	0.5673	0.3390
AutoInt [7]	0.6415	0.5919	0.3296
VBPR [12]	0.6714	0.6053	0.3239
MLFM [11]	0.6862	0.6063	0.3179
DCIM [10]	0.7604	0.6092	0.3026
CMBF	**0.7880**	**0.6118**	**0.3001**

4.4. Ablation Study

4.4.1. Influence of Parameters

In this section, we study the contribution of cross-modal fusion module to the final performance with ablation experiments on the MovieLens datasetand Amazon dataset. As shown in Table 6, the parameter N_c denotes the number of the cross-modal fusion layers stacked in the module, and 0 denotes that the cross-modal fusion module is not in use. We have the following observations from Table 6: first, with the addition of cross-modal fusion module, the recommendation performance improves at least 0.5% in terms of AUC and 0.48% in terms of GAUC on the MovieLens dataset (0.41% and 0.56% on the Amazon dataset). Second, when the N_c is not exceeding 3, the recommendation performance improves with the increase in N_c. The highest increase reaches 0.86% in terms of AUC and 0.87% in terms of GAUC on the MovieLens dataset (1.19% and 1.36% on Amazon dataset). Third, when the N_c is more than 3, the recommendation performance drops instead. The possible reason is that the dimension of the extracted features is too high to remain low-level features.

Table 6. Influence of parameter N_c on the performance of CMBF. The results in bold represent the best performances.

	MovieLens			Amazon		
N_c	AUC	GAUC	Logloss	AUC	GAUC	Logloss
0	0.8750	0.8276	0.3417	0.7761	0.5997	0.3123
1	0.8800	0.8324	0.3352	0.7802	0.6053	0.3046
2	0.8831	0.8361	0.3312	0.7848	0.6087	**0.2998**
3	**0.8836**	**0.8363**	**0.3302**	**0.7880**	**0.6118**	0.3001
4	0.8824	0.8357	0.3321	0.7875	0.6072	0.3038
5	0.8813	0.8327	0.3333	0.7866	0.6028	0.3095

4.4.2. Evaluation of Model Efficiency

We compare the run time per epoch of CMBF and other algorithms, as shown in Figures 4 and 5. Obviously, the run time of LR and FM is shortest because they do not use a deep network. The single-modal algorithms (i.e.; DeepFM, Wide&Deep, AutoInt) require less run time than multi-modal algorithms with worse performance of prediction results. Note that compared with other mult-modal algorithms (i.e.; VBPR, MLFM, DCIM), our CMBF does not require much run time, and its performance is improved, which proves the efficiency of CMBF.

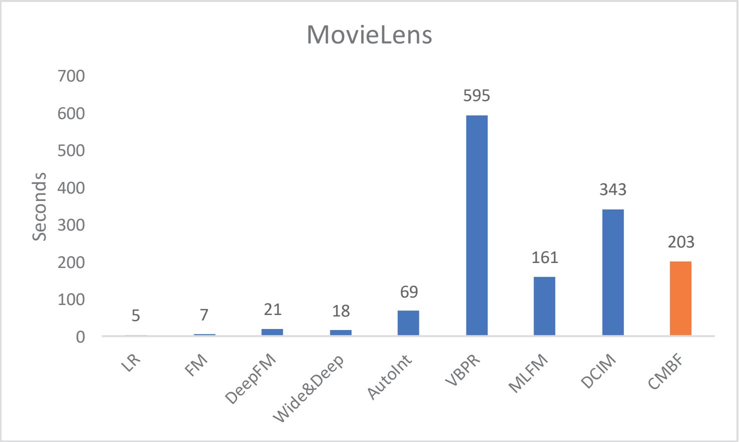

Figure 4. Run time per epoch of different algorithms on the MovieLens dataset.

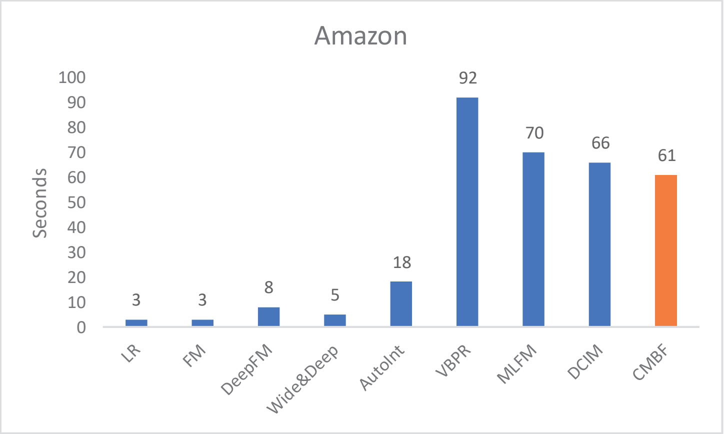

Figure 5. Run time per epoch of different algorithms on the Amazon dataset.

5. Conclusions

In this paper, we proposed a Cross-Modal-Based Fusion Recommendation Algorithm (CMBF), which can alleviate the data sparsity problem in the recommendation system. The key to our algorithm is mining the relevance between two modalities and trying to obtain the high-level feature representation containing more information. Compared to existing multi-modal algorithms that use the simple fusion method, we propose the cross-modal fusion method to completely fuse the multi-modal features. We conduct experiments on two datasets and compared with other algorithms. The experimental results show that our proposed CMBF achieves the best recommendation performance. In addition, the ablation study proves that our cross-modal fusion method is an innovation in the multi-modal fusion field. However, the algorithm proposed in this paper is only suitable for the fusion of two modal features, and how to expand to three or more modal features requires further consideration.

Author Contributions: Conceptualization, X.C. and Y.L.; methodology, X.C.; software, X.C.; validation, X.C.; Y.L. and Y.W.; formal analysis, Y.W. and J.Y.; investigation, X.C.; Y.L. and Y.W.; resources, J.Y.; data curation, X.C.; writing—original draft preparation, X.C.; writing—review and editing, Y.W. and J.Y.; visualization, X.C. and Y.L.; supervision, Y.W.; project administration, Y.W. and J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Adomavicius, G.; Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 734–749. [CrossRef]
2. Christakou, C.; Vrettos, S.; Stafylopatis, A. A hybrid movie recommender system based on neural networks. *Int. J. Artif. Intell. Tools* **2007**, *16*, 771–792. [CrossRef]
3. Salter, J.; Antonopoulos, N. CinemaScreen recommender agent: Combining collaborative and content-based filtering. *IEEE Intell. Syst.* **2006**, *21*, 35–41. [CrossRef]
4. Gunawardana, A.; Meek, C. Tied boltzmann machines for cold start recommendations. In Proceedings of the 2008 ACM Conference on Recommender Systems, Lausanne, Switzerland, 23–25 October 2008; pp. 19–26.
5. Shen, X.; Yi, B.; Zhang, Z.; Shu, J.; Liu, H. Automatic recommendation technology for learning resources with convolutional neural network. In Proceedings of the 2016 International Symposium on Educational Technology (ISET), Beijing, China, 19–21 July 2016; pp. 30–34.
6. Unger, M.; Bar, A.; Shapira, B.; Rokach, L. Towards latent context-aware recommendation systems. *Knowl. Based Syst.* **2016**, *104*, 165–178. [CrossRef]
7. Song, W.; Shi, C.; Xiao, Z.; Duan, Z.; Xu, Y.; Zhang, M.; Tang, J. Autoint: Automatic feature interaction learning via self-attentive neural networks. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 1161–1170.
8. Li, Z.; Cheng, W.; Chen, Y.; Chen, H.; Wang, W. Interpretable Click-Through Rate Prediction through Hierarchical Attention. In Proceedings of the 13th International Conference on Web Search and Data Mining, Houston, TX, USA, 3–7 February 2020; pp. 313–321.
9. Cai, J.J.; Tang, J.; Chen, Q.G.; Hu, Y.; Wang, X.; Huang, S.J. Multi-View Active Learning for Video Recommendation. In Proceedings of the IJCAI, Macao, China, 10–16 August 2019; pp. 2053–2059.
10. Ge, T.; Zhao, L.; Zhou, G.; Chen, K.; Liu, S.; Yi, H.; Hu, Z.; Liu, B.; Sun, P.; Liu, H.; et al. Image matters: Visually modeling user behaviors using advanced model server. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Turin, Italy, 22–26 October 2018; pp. 2087–2095.
11. Wu, C.; Wu, F.; An, M.; Huang, J.; Huang, Y.; Xie, X. Neural News Recommendation with Attentive Multi-View Learning. *arXiv* **2019**, arXiv:1907.05576.
12. 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; Volume 30, pp. 144–150.

13. Bourlard, H.; Dupont, S. A mew asr approach based on independent processing and recombination of partial frequency bands. In Proceedings of the 4th International Conference on Spoken Language Processing ICSLP'96, Philadelphia, PA, USA, 3–6 October 1996; Volume 1, pp. 426–429.
14. Valstar, M.; Schuller, B.; Smith, K.; Eyben, F.; Jiang, B.; Bilakhia, S.; Schnieder, S.; Cowie, R.; Pantic, M. Avec 2013: The continuous audio/visual emotion and depression recognition challenge. In Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, Barcelona, Spain, 21 October 2013; pp. 3–10.
15. Hodosh, M.; Young, P.; Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.* **2013**, *47*, 853–899. [CrossRef]
16. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
17. Chen, Q.; Wang, W.; Huang, K.; De, S.; Coenen, F. Multi-modal generative adversarial networks for traffic event detection in smart cities. *Expert Syst. Appl.* **2021**, *177*, 114939. [CrossRef]
18. Yang, B.; Mei, T.; Hua, X.S.; Yang, L.; Yang, S.Q.; Li, M. Online video recommendation based on multimodal fusion and relevance feedback. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 9–11 July 2007; pp. 73–80.
19. Oramas, S.; Nieto, O.; Sordo, M.; Serra, X. A deep multimodal approach for cold-start music recommendation. In Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems, Como, Italy, 27 August 2017; pp. 32–37.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
21. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef] [PubMed]
22. Harper, F.M.; Konstan, J.A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst. (TIIS)* **2015**, *5*, 1–19. [CrossRef]
23. Ni, J.; Li, J.; McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 5–7 December 2019; pp. 188–197.
24. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
25. McMahan, H.B.; Holt, G.; Sculley, D.; Young, M.; Ebner, D.; Grady, J.; Nie, L.; Phillips, T.; Davydov, E.; Golovin, D.; et al. Ad click prediction: A view from the trenches. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; pp. 1222–1230.
26. Rendle, S. Factorization machines. In Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia, 13–17 December 2010; pp. 995–1000.
27. Guo, H.; Tang, R.; Ye, Y.; Li, Z.; He, X. DeepFM: A factorization-machine based neural network for CTR prediction. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 1725–1731.
28. Cheng, H.T.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; Ispir, M.; et al. Wide & deep learning for recommender systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, Boston, MA, USA, 15 September 2016.
29. Wang, Q.; Huang, P.; Xing, S.; Zhao, X. Click-Through Rate Prediction Combining Mutual Information Feature Weighting and Feature Interaction. *IEEE Access* **2020**, *8*, 207216–207225. [CrossRef]
30. Wang, Q.; Huang, P.; Xing, S.; Zhao, X. A Hierarchical Attention Model for CTR Prediction Based on User Interest. *IEEE Syst. J.* **2020**, *14*, 4015–4024. [CrossRef]
31. Kingma, D.P.; Ba, J. Adam: Method for stochastic optimization. *arXiv* **2015**, arXiv:1412.6980.
32. Zhu, H.; Jin, J.; Tan, C.; Pan, F.; Zeng, Y.; Li, H.; Gai, K. Optimized cost per click in taobao display advertising. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 2191–2200.