MULTI-MODAL REPRESENTATION LEARNING FOR SHORT VIDEO UNDERSTANDING AND RECOMMENDATION

[1]*Daya Guo,* [2]*Jiangshui Hong,* [3]*Binli Luo,* [4]*Qirui Yan, and* [3,5]*Zhangming Niu*

[1]Sun Yat-Sen University [2]Durham University, and Simula Research Laboratory
[3]MindRank AI [4]South China University of Technology
[5]Aldaddin Healthcare Technologies SE

guody5@mail2.sysu.edu.cn jiangshui.hong@durham.ac.uk
binli@mindrank.ai eeqryan@mail.scut.edu.cn zhangming@aladdinid.com

ABSTRACT

We study the task of short video understanding and recommendation which predicts the user's preference based on multimodal contents, including visual features, text features, audio features and user interactive history. In this paper, we present a multi-modal representation learning method to improve the performance of recommender systems. The method first converts multi-modal contents into vectors in the embedding space, and then concatenates these vectors as the input of a multi-layer perceptron to make prediction. We also propose a novel Key-Value Memory to map dense real-values into vectors, which could obtain more sufficient semantic in a nonlinear manner. Experimental results show that our representation significantly improves several baselines and achieves the superior performance on the dataset of ICME 2019 Short Video Understanding and Recommendation Challenge.

***Index Terms*—** Multi-modal Representation, Factorization Machine, Key-Value Memory, Word2Vec, DeepWalk

1. INTRODUCTION

Short video understanding and recommendation aims to model the user's interest through video and user interaction history in order to predict the user's click behavior, i.e. the prediction of click-through rate (CTR). More recently, the neural network based approaches [1, 2, 3, 4, 5] have achieved promising performance in CTR prediction. These methods mainly extract categorical feature interactions or utilize graph embeddings for CTR prediction. Meanwhile, other existing works tend to mainly focus on single-modal features. Without consideration of multi-modal features, such as multi-media, user click history, and social graph in the field of video content understanding, the performance of the recommender system will be negatively affected.

In this work, we study how to make use of multi-modal features with a different structure. Specifically, our approach first converts multi-modal contents to vectors in the embedding space, and then concatenates these vectors as the input of multi-layer perceptron (MLP). Meanwhile, we propose a novel Key-Value Memory [6] to map dense real-values into vectors, which could obtain more sufficient semantic in an nonlinear manner. One notable drawback of training MLP is that the distribution of embeddings varies with different contents, which adversely slows down the training. To address this problem, we apply batch normalization to the input.

We conduct experiments on the dataset of ICME 2019 Short Video Understanding and Recommendation Challenge. This challenge provides multi-modal video features, including visual features, text features, audio features, and user interactive history. According to this user interactive history, we are now able to construct social graphs between users. The task aims to predict the probability that each user will finish watching and like a given video. Results show that by applying our multi-modal representation to recommender systems, it yields further improvements with a 1.5% gain, clearly demonstrating the effectiveness of the proposed multi-modal representation learning.

2. RELATED WORK

Personalized recommendation is a fundamental task in machine learning that learns the ability to predict which items (i.e. videos in our work) will be considered interesting by the user. Recently, the dimension and form of these features have become larger and more diverse. At the same time, the structure of existing models has evolved from shallow to deep.

Linear models, such as logical regression with FTRL [7], are widely adopted as they are easy to manage, maintain, and deploy. However, linear models lack the ability of learning hidden features, which requires significant engineering cross features in order to achieve a better performance. Therefore, some researchers exploit boosting decision trees [8] to help build feature transformations [9]. In order to overcome the problem of high-dimensional and sparse features, factoriza-

Jiangshui Hong is the corresponding author

978-1-5386-9214-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMEW.2019.00134

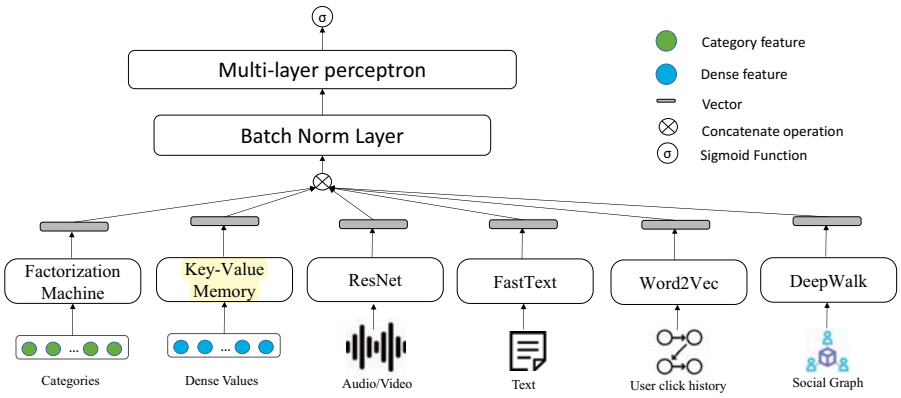

Fig. 1. An overview of our approach that converts multi-modal contents into embeddings to make prediction.

tion machines [10, 11] embedding each feature into a low dimension latent vector. Compared with non-factorization models, they are able to learn feature interactions because recommendations are made via the product of two latent vectors.

Recently, neural models have proven effective in recommendation systems [1, 2, 3, 4, 5]. These methods attempt to automatically learn patterns from categorical feature interactions. Differing from these works, we consider multi-modal features with more complex structure, such as multi-media, user history sequence and social graph. Our approach converts multi-modal contents to the same fashion (i.e. embedding) to boost recommender systems.

3. METHODOLOGY

We will first provide an overview of our approach, and then describe a multi-modal representation learning method which convert various types of contents into latent vectors in the embedding space.

3.1. Overview of the Approach

Figure 1 gives an overview of our model. First, given multi-modal contents, we utilize various methods to convert these contents into embeddings. Since distribution of embeddings varies with different contents, which adversely slows down the training, a batch norm layer is used to normalize these vectors. Lastly, the concatenation of these vectors is viewed as the input to predict whether each user will finish watching and like a given video by MLP following a sigmoid function.

3.2. Multi-Modal Representation

As illustrated in Figure 1, the task includes six types of features, including categories, dense real-values, audio/video, text, user click history, and social graph. We describe these features and corresponding methods to convert them into vectors one after another.

Category Features Category features are highly sparse but helpful for short video understanding and recommendation [10]. For examples, users usually follow some specific creator's videos, which could help us to predict whether users will like these videos. In order to extract interactive features among categories with huge sparsity, we utilize Deep Factorization Machine to obtain features interactions. Specifically, category features are first fed into an embedding layer to obtain corresponding embeddings, and then a compressed interaction network (CIN) [3] is used to generate feature interactions in an explicit fashion and at the vector-wise level.

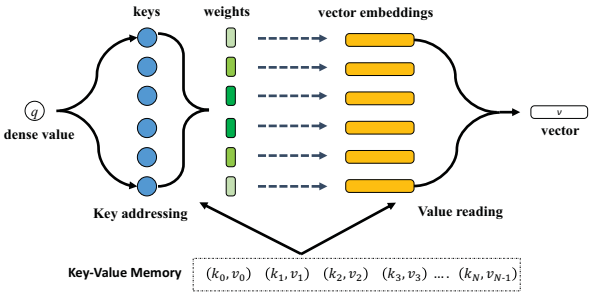

Fig. 2. An overview of Key-Value Memory that consists of key addressing and value reading.

Dense Real-values Traditional methods to use dense real-values are directly fed into a neural network or split into several buckets as category features. Instead, we propose a novel method to utilize key-value memory [6] to map dense real-values into vectors. The basic idea is to weighed averaging trainable vector embeddings. As shown in Figure 2, the key-value memory network consists of two stages. At the first stage (i.e. key addressing), we calculate weights w_i of trainable vector embeddings v_i given a dense value q as Equation 1. At the second stage (i.e. value reading), we obtain the final representation $v = \sum_{i=0}^{N-1} w_i v_i$ of the dense real-value according to the weights so far.

$$w_i = softmax(\frac{1}{|q - k_i| + \epsilon}) \quad (1)$$

where N is the number of trainable vector embeddings, k_i is the key of the i-th vector v_i, and the hyper parameter ϵ is tuned on the dev set. We normalize the dense value q ranges from 0 to 1 and set k_i as $\frac{2i+1}{2N}$.

Multi-media Multi-media contents in short videos recommendation usually include videos, audios and texts (e.g. tiles), which provides high level of information concerning recommended short videos. We adopt Resnet [12] to encode videos and audios into embeddings. In addition, we use FastText [13] to encode text by averaging word embeddings.

User Click History As users usually tend to follow similar authors or short videos, we can subsequently cluster them. We describe how to cluster short videos and obtain embeddings of them. The calculation of author embeddings is analogous to short video embeddings. A click history of the user u is a sequence of videos $[x_0^u, x_1^u, ..., x_n^u]$, where i-th video x_i^u is clicked after clicking x_{i-1}^u by the user. We can treat the sequence as a sentence and videos as words. And then we utilize Word2Vec [14] to represent videos, which can cluster similar short videos in the embedding space. Finally, we extract the embedding of current video clicked by the user as history features to make final prediction.

Social Graph Social graphs can reflect social relations between users, which facilitates the prediction of user preference through statistical models. We define a user-video social graph, including two kinds of vertices (e.g. users and videos respectively). The edge from a user vertex to a video vertex means that the user clicked the video once. We then adopt DeepWalk [15], a approach for learning latent representations of vertices, on this social graph to obtain latent vectors of vertices. These latent vectors encode social relations in the embedding space, which can be fed into the neural network. Besides, we also define several social graph to improve our model, such as a user-author social graph.

3.3. Learning

In this section, we describe how to learn our multi-modal representation. With regard to Word2Vec and DeepWalk, these methods can learn representations of user click history and social graph in an unsupervised setting (more detail about learning could be found in [14, 15]). Representations of audio and video are extracted from pre-train ResNet. Other embeddings will be trained in a supervised manner by maximizing the sum of log probabilities of ground truth.

Here, we list our training details. We set the dimension of Factorization Machine and FastText as 16, and the dimension of Word2Vec and Deepwalk as 64. The challenge has provided pre-train ResNet embedding. In the Key-Value memory, the number of trainable embeddings for each field is 100, dimension is 16 and hyper parameter ϵ is e^{-15}. Model parameters are initialized with uniform distribution and updated with Adam optimizer. We set the learning rate as 0.0002 and the batch size as 4096. The model is trained for only one epoch.

4. EXPERIMENT

We conduct experiments on the daset of ICME 2019 Short Video Understanding and Recommendation Challenge. The dataset is provided by the ByteDance Inc, including 19M/2M examples for training/testing for track 2. Each example consists of categorical features (e.g author and user information) and multi-media contents of short videos. In this competition, the task aims to predict the probability of finishing watching and liking a short video when a user browse a video. AUC (area under ROC curve) is used as the evaluation metric. The final score is obtained by weighted averaging finish and like scores with 0.7 and 0.3.

4.1. Model Comparisons

Methods	Score
FM [10] + Basic Features	74.81%
XdeepFM [3] + Basic Features	77.32%
Xgboost [8] + Basic Features	78.55%
Xgboost + Multi-modal Representation	79.64%
MLP + Multi-modal Representation	80.00%

Table 1. Performance of different approaches on the dataset.

We report the results of existing methods on the dataset and demonstrate that our multi-modal representation learning is an effective way to improve the performance of current recommender systems. We implemented several methods, including **FM** [10], which is a factorization machine; **XdeepFM**, a state-of-the-art wide&deep neural network on CTR prediction; and **Xgboost** which is a scalable tree boosting system. The underlying **Basic Features** allow models to only use categorical features to make a prediction, while **Multi-modal Features** is our multi-modal representation proposed in this paper. According to Table1, we can see that applying multi-modal representation to the Xgboost model yields further improvements with a 1% absolute gain. Furthermore, we use **MLP** with a batch norm layer by feeding our multi-modal representation to achieve superior performance with 80.00% scores, which demonstrates that our multi-modal representation is useful for different models.

4.2. Influences of Different Representation

We conduct ablation analysis to better understand how various components impact the overall performance in our representation. We remove each component to analyze their contribution, respectively. Table 2 shows that the score drops greatly from 80.00% to 79.03% when ablating DeepWalk, which

Methods	Score
Our approach − Factorization Machine	79.85%
Our approach − Key-Value Memory	79.86%
Our approach − ResNet	79.33%
Our approach − FastText	79.96%
Our approach − Word2Vec	79.90%
Our approach − DeepWalk	79.03%

Table 2. Performance of different representation.

reveals the importance of the social graph in the short videos recommendation scenario. After removing ResNet, the score drops to 79.33%. This is consistent with our intuition that the content of videos/audios are critical factors of recommending further short videos for users. We can see that other representations also bring 0.1%∼0.2% improvement.

5. CONCLUSION

In this paper, we present a multi-modal representation learning method which converts multi-modal contents into embeddings, and a novel method to map dense real-values to vectors by a Key-Value Memory. We show that our multi-modal representation learning method can further improve recommender systems and our proposed key-value memory could better utilize dense real-values.

6. REFERENCES

[1] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He, "Deepfm: a factorization-machine based neural network for ctr prediction," *arXiv preprint arXiv:1703.04247*, 2017.

[2] Xiangnan He and Tat-Seng Chua, "Neural factorization machines for sparse predictive analytics," in *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2017, pp. 355–364.

[3] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun, "xdeepfm: Combining explicit and implicit feature interactions for recommender systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1754–1763.

[4] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai, "Deep interest network for click-through rate prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1059–1068.

[5] Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee, "Billion-scale commodity embedding for e-commerce recommendation in alibaba," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 839–848.

[6] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston, "Key-value memory networks for directly reading documents," *arXiv preprint arXiv:1606.03126*, 2016.

[7] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, 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*. ACM, 2013, pp. 1222–1230.

[8] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[9] Xiaoliang Ling, Weiwei Deng, Chen Gu, Hucheng Zhou, Cui Li, and Feng Sun, "Model ensemble for click prediction in bing search ads," in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2017, pp. 689–698.

[10] Steffen Rendle, "Factorization machines," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 995–1000.

[11] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin, "Field-aware factorization machines for ctr prediction," in *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 2016, pp. 43–50.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[15] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.