iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35350

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Feature Re-Learning for Video Recommendation

Chanjal C
RIT Pampady Kottayam

Abstract: ***Predicting the relevance between two given videos with respect to their visual content is a key component for content-based video recommendation and retrieval. The application is in video recommendation, video annotation, Category or near-duplicate video retrieval, video copy detection and so on. In order to estimate video relevance previous works utilize textual content of videos and lead to poor performance. The proposed method is feature re-learning for video relevance prediction. This work focus on the visual contents to predict the relevance between two videos. A given feature is projected into a new space by an affine transformation. Different from previous works this use a standard triplet ranking loss that optimize the projection process by a novel negative-enhanced triplet ranking loss. In order to generate more training data, propose a data augmentation strategy which works directly on video features. The multi-level augmentation strategy works for video features, which benefits the feature relearning. The proposed augmentation strategy can be flexibly used for frame-level or video-level features. The loss function that consider the absolute similarity of positive pairs and supervise the feature re-learning process and a new formula for video relevance computation.***

Keywords: Feature Re-learning, Ranking Loss, Data Augmentation, Loss function, Feature space.

I. INTRODUCTION

Predicting the relevance between two given videos with respect to their visual content is a key component for content-based video recommendation and retrieval. In the case of video recommendation, the recommendation system suggest videos which may be of interest to a specific user. To that end, the video relevance shall reflect the user's feedback like watch, search and browsing history. For category video retrieval, one wants to search for videos that are semantically similar to a given query video, here the video relevance should reflect the semantical similarity. For near-duplicate video retrieval, one want to retrieve videos showing exactly the same story but with minor photographic differences and editions with respect to a given query video. For this purpose, the video relevance computed based on visual similarity. It is clear that the optimal approach to video relevance prediction is task dependent. In order to estimate video relevance, some works utilize textual content of videos. For instance, someone utilize metadata associated with videos, such as title, keywords, and directors. However, the metadata is not always available and its quality is not guaranteed, especially for user-generated videos. For example, the video title is easily alterable, which may be deliberately titled to attract users while irrelevant to the video content itself. Hence, video relevance prediction depending on the textual content lead to poor performance. In contrast to the textual content of videos, visual contents are available right after the video has created and more reliable. In this work focus on the visual contents to predict the relevance between two videos. The initial efforts for learning a new video feature space and then measuring the video relevance in the new space. This process transforms a given feature into a new space, this coin the feature re-learning. The essential difference between feature re-learning and traditional feature transform is twofold. The prefix "re" emphasizes the given feature is a learned representation. By contrast, the given feature in a traditional setting is low-level, e.g., bag of local descriptors. Improving an already learned feature is more challenging. Supervised learning is a must for feature re-learning. By contrast, traditional feature transformation can be unsupervised, e.g., Principle Component Analysis or random projection. For learning based methods, the choice of the loss function is important. The previous works utilize a triplet ranking loss which preserves the relative similarity among videos. The loss needs relevant video pairs for training and aims to make the similarity between relevant video pairs larger than that between irrelevant pairs in the learned feature space. Note that the triplet ranking loss ignores how close (or how far) between the relevant (or irrelevant) video pairs, which affects its effectiveness for training a good model for video relevance prediction. In this work, propose a novel negative-enhanced triplet ranking loss (NETRL). The new loss effectively considers both relative and absolute similarity among videos. In this work study the video relevance prediction in the context of the Hulu Content based Video Relevance Prediction Challenge. In this challenge participants want to recommend a list of relevant videos with respect to the given seed video from a set of candidate videos here only given a seed video without any metadata. The key of the challenge is to predict relevance between a seed video and a candidate video. Notice that the participants have no access to original video data. Instead, the organizers provide two visual features, extracted from individual frames and frame sequences.

A version of this work was published [11]. In this work improve the feature re-learning process by dimensionality reduction and improve the loss function by considering the absolute similarity among the relevant pairs. The paper is organized as follows. Chapter 1 provides an introduction to the system. Chapter 2 describes the existing system with adetailed study. Chapter 3 describes the system design, which includes the proposed system and the methodology. Chapter 4 provides the evaluation of the model and chapter 5 concludes the paper.

II. RELATED WORKS

Video relevance prediction between two videos with respect to their visual content feature is important for content based video recommendation and their retrieval. The current system performs video relevance prediction using original content of the video and that lead to the copyright privacy issue of the content. A. Karpathy, G. Toderici et al, proposed ”Large-scale Video Classification with Convolutional Neural Networks”[1]. Treat every video as short fixed sized clips. The videos varies in temporal extent so cannot processed using fixed architecture. The different connectivity patterns like early fusion, late fusion and slow fusion used. The frame feature of the video is considered for the classification operation. The frame feature is fed into two spatial resolution processing called context stream and fovea stream. The processing consists of convolution neural network steps like convolution, normalization and pooling. This method handles large scale data using two separate spatial resolution stream and thus improve the performance. The different resolution streams also increase the complexity of the system. J. Lee, S. Abu-El-Haija et al, proposed ”Large Scale Content-Only Video Recommendation” [2], using the content of the video. This is a video content based similarity learning problem predict the relationship between the videos. They embedded all the videos into an embedding space and the similar videos located closest to each other. Both the audio and visual features are used as input. The similar video is identified based on the co-watched system. The feed forward network is designed and concatenated the audio and visual features to identify similarity. They perform recommendation including new video uploads that improve the performance. The co-watched system similarity calculation does not provide accurate result it may contain irrelevant video pairs.

Y. Bhalgat, A. Arbor et al, proposed ”Fused LSTM at ACMMM-2018 CBVRP Challenge: Fusing frame-level and video-level features for Content-based Video Relevance Prediction”[3]. The frame level and video level feature extracted using the inception pool and C3D pool methods. Using triplet loss function, generate triplets of similar and dissimilar videos. The preextracted features of each triplet passed fused LSTM and concatenate the features. Then calculate the similarity using kernel based similarity function. This system considers both frame level and video level features so this will improve the performance of the relevance prediction system. The feature extraction of each triplet makes an additional overhead and affects the system performance. X. Du, H. Yin et al, proposed ”Personalized Video Recommendation Using Rich Contents from Videos”[4]. The system used the method of collaborative embedding regression for the recommendation. Use input as the rich content of the videos and also user rating matrix. The user feedback is converted to the rating matrix. The CER module creates user-video pairs for each user and from that generate top-k recommendation. The system work well for the videos there specific content feature is unavailable. The similarity of video identified based on the user feedback and rich content this may contain irrelevant data and the efficiency of the system may affected.

J. Dong, X. Li et al, proposed ”Feature Re-Learning with Data Augmentation for Contentbased Video Recommendation ”[5]. The system provides the recommendation for the privacy protected videos. They perform frame level and video level augmentation to create sufficient data. Traditional augmentation not possible in privacy protected videos. The feature relearning method improves the efficiency of the system. The frame level and video level feature obtained is mapped to a new feature space and designed it as one layer fully connected network. The system not consider the similarity among relevant pairs and the number of features induces overhead . M. Liu, X. Xie et al, proposed”Content-based Video Relevance Prediction Challenge: Data, Protocol, and Baseline”[6]. This system generated similar videos without using the original content of the videos instead they use the deep neural network. features for the video representation. Specifically frame-level and video-level features are used. These off the shelf features are mapped to feature space and train the model using the triplet ranking loss function. Initially create triplets from the video set. The anchor video randomly selected from the training data, the positive video samples from the relevant video pair. The proposed method not used the original content thus preserve the privacy of the data. Here off the shelf visual features are directly used so this will decrease the accuracy.

J. Dong, X. Li et al, proposed ”Hybrid Space Learning for Language-based Video Retrieval”[7]. The proposed system will retrieve videos based on the user ad-hoc queries. The network encodes the inputs such as query sentence or video and senses it in parallel. The video encoded using the mean pooling technique. For each video, per frame extract deep features using the convolutional neural network and converted to words. The model is trained using the word2vec architecture and then mapped to latent space and concept space and find the cosine similarity.

The complexity of the system is lower due to the parallel processing. The system used the original video feature for the similarity calculation. J. Dong, X. Li et al, proposed ”Predicting Visual Features From Text for Image and Video Caption Retrieval”[8]. The system solves the caption retrieval problem in visual space and finds the best sentence that describe the content of the image and video. The proposed system not used the original content and preserves data security issues. The method used here is word2visualvec architecture. This is a deep neural network architecture that learns to predict a visual feature representation from textual input. Initially perform a multi-scale sentence vectorization to handle the varying length sentences. This utilizes BoW, word2vec and RNN based text encodings. The objective function used here is mean squared error. Train the Word2VisualVec to minimize the overall MSE loss on a given training set it contains a number of relevant image-sentence pairs. For video caption retrieval, project sentences into the video feature space. The accuracy of the system is lower due to the misleading texts.

J. Dong, X. Li et al, proposed”Dual Encoding for ZeroExample Video Retrieval”[9]. Here retrieve videos by ad-hoc queries described in natural language text. This network that encodes the input like a query sentence or a video. The proposed dual encoding network that encode sentence and video in a dual manner. For video extract sequence of frames from the videos. Then perframe extract deep video features using CNN. The system used the common space learning and these representations can be transformed to perform sequenceto-sequence cross-modal matching effectively. Once the network is trained, encoding at each side is performed independently and process large-scale videos offline and answer ad-hoc queries on the fly. The system uses the original content of the video and system performance is low. L. Jing, Y. Tian et al, proposed ”Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey”[10]. The proposed system is a ranking oriented visual feature re-learning method. Here summarize the pretext tasks into four categories. Generation based Method learn visual features by solving pretext tasks that involve image or video generation. The Context-based pretext tasks mainly employ the context features of images or videos and the Free Semantic Label-based Method train networks with automatically generated semantic labels. This work not find out how far relevant or irrelevant the features.

III. METHOD

First introduces a ranking-oriented feature re-learning method which maps an off-the-shelf video feature into a new feature space, followed by multi-level augmentation strategy which considers both the frame-level and video-level features. Where the higher dimension reduced to a lower dimension .Finally present a strategy to predict video relevance in the re-learned video feature space.

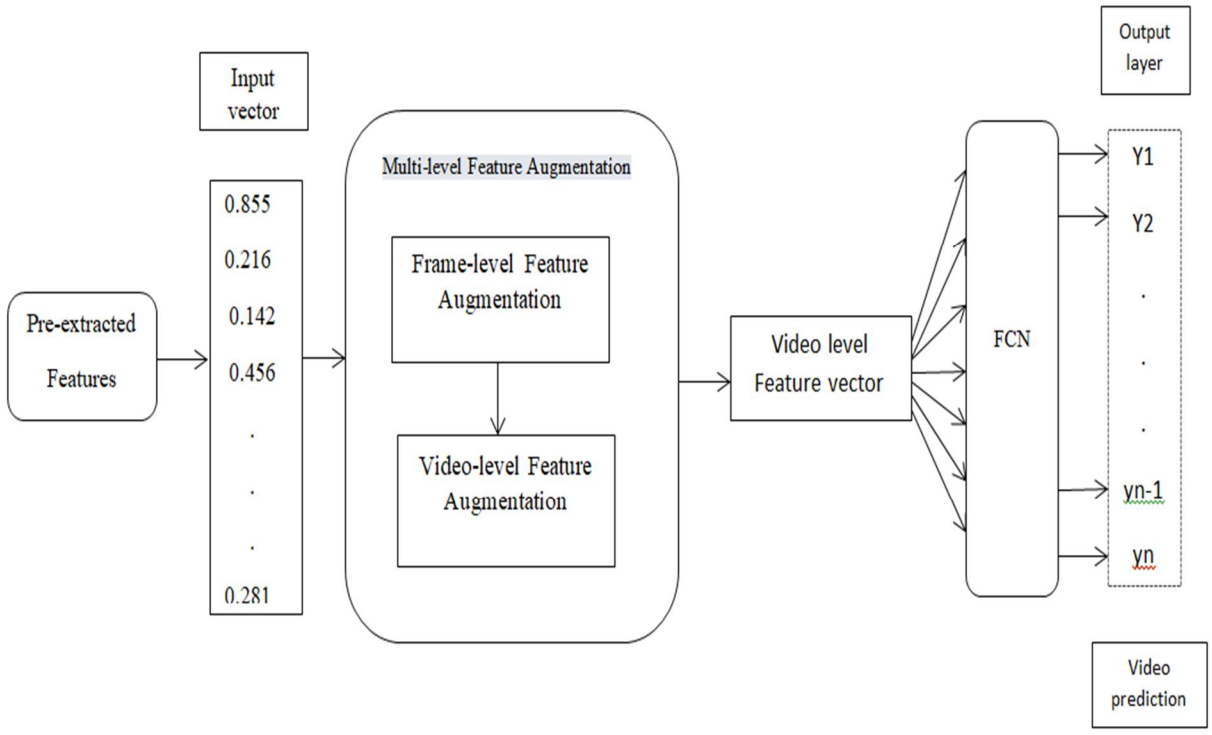


Fig 1. Overall Design

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

A. Multi-level Feature Augmentation

The performance of deep neural network decreased when the amount of training samples less. Data augmentation is a method to increase the data diversity without collecting new training data. In the case of privacy protected videos, the original video is not given and only the pre-computed features[1] will be available and in this case the traditional data augmentation method is not possible. The multi-level feature augmentation [11] method increase the training samples by performing augmentation on video features without original videos. This augmentation method has 2 steps: frame-level feature augmentation and video-level feature augmentation. Figure 1 describes the multilevel feature augmentation.

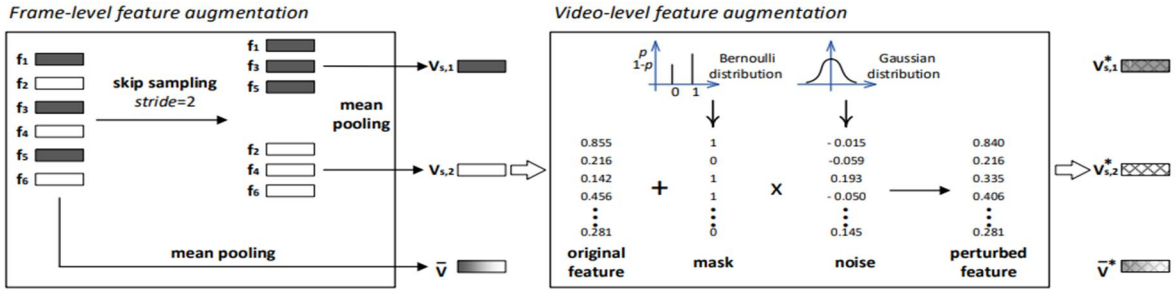


Fig 2. Multi-level augmentation strategy for features. It consists of two steps: frame-level feature augmentation and video-level feature augmentation.

1) *Frame Level Augmentation:* This augmentation generate more training samples by employing mean pooling over the frame sequences. Initially perform a skip sampling with a stride value s over the frame sequence and further employ mean pooling over the set of frame sequences and generate s new frame sequence. Finally employ a mean pooling over the entire frame sequence and generate s+1 video level features.
2) *Video level Augmentation***:** Output of frame level augmentation is the input for video level augmentation. This augmentation strategy removes the noise generated during the feature extraction. In video level augmentation a perturbed video level feature is generated by using a mask with Bernoulli's random variable with the probability of being 1 is 0.5 and randomly generated Gaussian noise.

B. Feature Re-learning

Feature re-learning is a method to map the off-the-shelf video features to a new feature space. This projection is designed as a fully connected network. In the new feature space the relevant videos are near and irrelevant videos are far away. Before feeding the videos to feature re-learning model first represent each video as a video level feature vector. The re-learning model project the video features to a new feature space with a reduced dimensionality.

The model can be trained using negative enhanced triplet ranking loss function[11]. Here firstly describe the triplet ranking loss function(TRL). TRL is widely used in many ranking based task[5],[6]. The loss function needs triplets for the training, here construct a large set of triplets T = {v,v+,v-}, from the relevant video pair set, where v+ and v- indicate the relevant and irrelevant video with respect to video v. Given a triplet of (v, v+, v−), the TRL for the given triplet is defined as follows:

$$L\ (v, v+, v-) = \max(0, m1 - cs\varphi(v, v+) + cs\varphi(v, v-))$$

Where csφ(v, v+) indicate the cosine similarity score between the video v and its relevant video v+ and csφ(v, v−) indicate the cosine similarity score between the video v and irrelevant video v-. The irrelevant video is randomly chosen from the training set. The TRL only consider the relative similarity score between the triplets they ignore how much close the relevant pair video set and how much far the irrelevant video pair set. The NETRL loss function which consider both the relative and absolute similarity between the video pairs. Here add a negative constraint max(0, csφ(v, v−) − m2), which enforce the similarity of negative video pair smaller than a given constant m2. When the similarity of negative pair smaller than the given constant m2 , the constraint will adjust the model to push negative video pair far away in the re-learned feature space. The NETRL computed as,

$$L(v, v+, v-; W, b) = \max(0, m1 - cs\varphi(v, v+) + cs\varphi(v, v-)) + \alpha \max(0, cs\varphi(v, v-) - m2),$$

Finally train the re-learning model by minimizing the proposed NETRL loss function on the triplet set.

The NETRL loss function only considering the absolute similarity between the negative pair, computing the absolute similarity between positive pair also make advantage for video relevance prediction. The constraint max(0, m1-csφ(v, v−)) will consider positive pair similarity. Also the dimensionality reduction will improve the system accuracy.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

C. Fully Connected Network

Fully connected neural network are type of artificial neural network, where all the neurons or node in one layer are connected to the next layer and each connection has its own weight. The advantage of fully connected network is that they are structure agnostic that is there are no special assumptions needed to be made about the input.

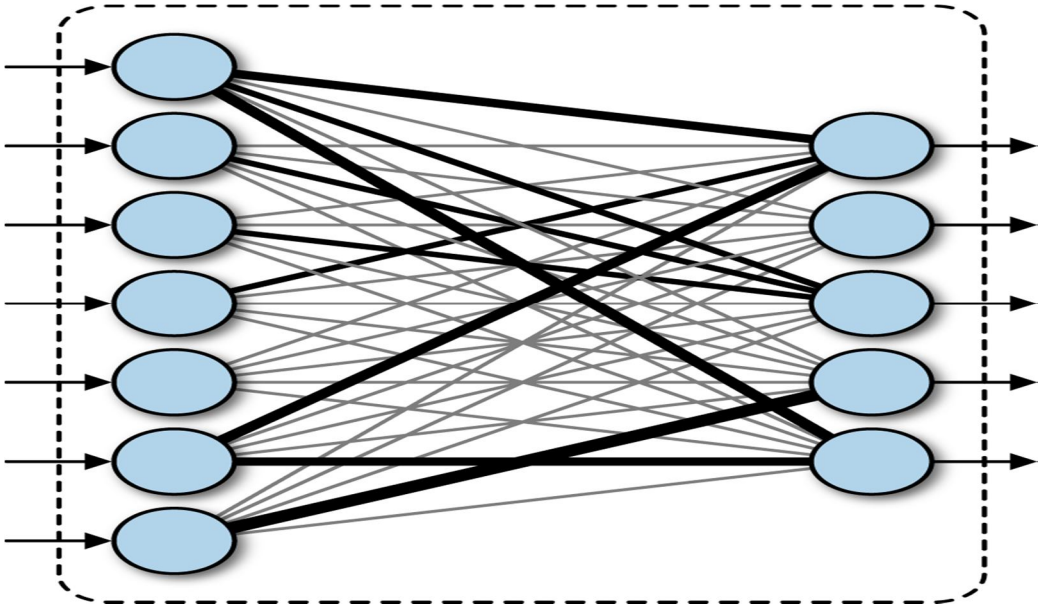


Fig 3. Fully Connected Network

D. Video Recommendation

In the context of the Hulu Content-based Video Relevance Prediction Challenge[6], the key is to predict relevance between a seed video and a candidate video. Depending on whether a candidate video is known to be relevant to another candidate video the relationship of a candidate video to another candidate video is unknown. Consequently, one has to fully count on the provided video features to predict the relevance between a given seed video and a candidate video. In this estimate their video relevance by the cosine similarity in the re-learned video space. The relationship of a candidate video to another candidate video is unknown then the relevant score can be computed as,

$$r(vs, vc) = cs\varphi(vs, vc).$$

Given a set of candidate videos, here sort the candidate videos in the descending of their relevance score with respect to a given seed video and consequently recommend the top k videos.

IV. EVALUATION

In order to verify the performance of the proposed system use the TV-shows dataset provided by HULU in the context of the Content-based Video Relevance Prediction Challenge[6]. The dataset has been divided into three subsets for training, validation and test. Detailed data split is as follows: training / validation / test of 3,000 / 864 / 3,000 videos for the TV-shows dataset. For each video in the training and validation set, it is associated with a list of relevant videos derived from implicit viewer feedbacks. The HULU challenge doesn't provide original videos while they provide pre-computed video features that are frame-level features. Frame level features generated using inception V3 networks trained on ImageNet dataset.

A. Performance Metric

The performance metric recall@k and hit@k are used for the evaluation of the proposed method. The metric recall is the fraction of relevant instances that were retrieved and hit is a user tend to browse top ranked videos in the first few pages, so hit with smaller k reflect model effectiveness. Here the k value indicate the top-k recommendation.

Figure 3 shows the performance curves of feature re-learning with varied dimensionality of the new video feature space on the TV-shows. Here compare the results of the dimensionality in the range of 32, 258, 512, 1024 and 2048 on the datasets and the best overall performance is reached with the dimensionality of 512, while the too small or too large dimensionality degrades the performance. So here set the dimensionality of the feature space as 512.

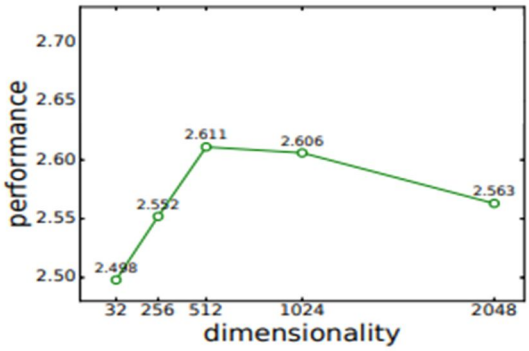


Fig.4 performance curves of feature re-learning with varied dimensionality of the new video feature space on the TV-shows

In order evaluate the performance of the proposed loss function compare it with commonly used ranking loss function TRL. The performance comparison is shows in Table 1. TRL consider only the relative similarity and NETRL loss considers both absolute and relative similarities on the datasets.

TABLE 4
Performance comparison of feature re-learning with different loss on the validation

Dataset	Loss Function	hit@k			recall@k		
TV-shows	TRL	0.333	0.440	0.503	0.193	0. 290	0.354
	NETRL	0.391	0.483	0.539	0.208	0.299	0.365

V. CONCLUSION

To predict task-specific video relevance proposes a ranking-oriented feature re-learning model with feature level data augmentation. The proposed multi-level data augmentation improves the performance of learning based models especially when the training data are inadequate. Ranking-oriented feature re-learning method to map videos into a new feature space where relevant videos are near and irrelevant videos are far away. Finally predict relevance between a seed video and a candidate video. The video relevance estimated by the cosine similarity in the re-learned video space.

REFERENCES

[1] A. Karpathy, G. Toderici et.al, “Large-scale video classification with convolutional neural networks,” in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 1725–1732.

[2] J. Lee and S. Abu-El-Haija et.al, “Large-scale content-only video recommendation,” in Proceedings of the IEEE International Conference on Computer Vision Workshop, 2017, pp. 987–995. [3]

[3] Y. Bhalgat et.al, “Fusedlstm: Fusing frame-level and video-level features for contentbased video relevance prediction,” arXiv preprint arXiv:1810.00136, 2018.

[4] X. Du, H. Yin et.al, “Personalized video recommendation using rich contents from videos,” IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 3, 2020.

[5] J. Dong, X. Li2 et.al, “Feature Re-Learning with Data Augmentation for Contentbased Video Recommendation,” in Proceedings of the ACM International Conference on Multimedia, 2018, pp. 2058–2062.

[6] M. Liu, X. Xie et.al, “Content-based video relevance prediction challenge: Data, protocol, and baseline,” arXiv preprint arXiv:1806.00737, 2018.

[7] J. Dong, Xirong Li et.al, “Hybrid Space Learning for Language-based Video Retrieval,” IEEE Transactions on Knowledge and Data Engineering, vol. 39, no. 12, pp. 2423–2436, 2020.

[8] J. Dong, X. Li et.al, “Predicting Visual Features From Text for Image and Video Caption Retrieval,” IEEE Transactions on Multimedia, vol. 20, no. 12, 2018.

[9] J. Dong, X. Li et.al, “Dual Encoding for Zero-Example Video Retrieval” IEEE conference on Computer Vision and Pattern Recognition, 2019, pp. 9346–9355.

[10] L. Jing, Y. Tian et.al, "Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 3, 2019.

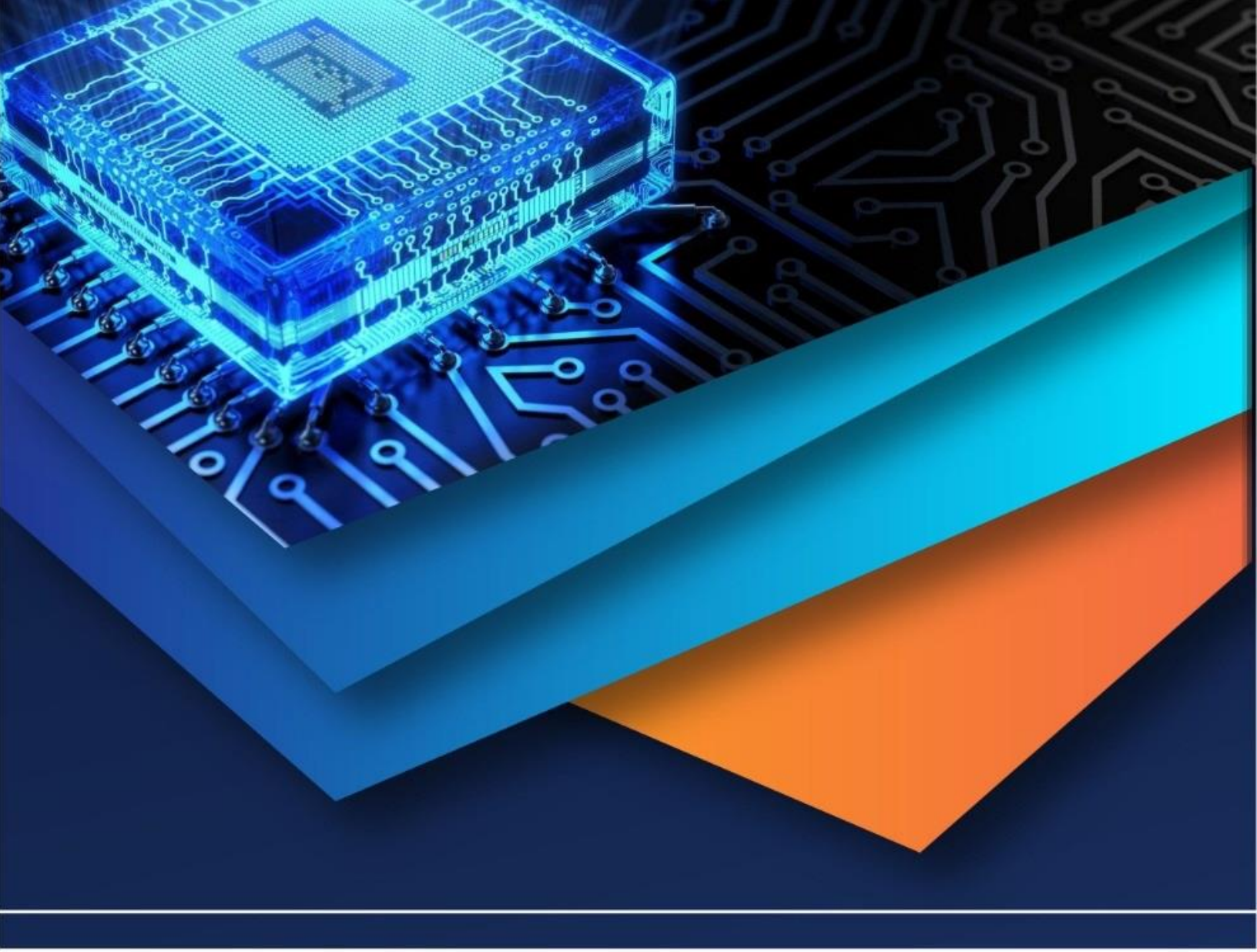
[11] J Dong, X. Wang et.al "Feature Re-Learning with Data Augmentation for Video Relevance Prediction," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 12, 2019.

[12] X. He, Z. He et.al, "Nais: Neural attentive item similarity model for recommendation," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 12, pp. 2354–2366, 2018.

[13] X. Liu, L. Zhao et.al, "Deep hashing with category mask for fast video retrieval," arXiv preprint arXiv:1712.08315, 2017.

[14] J. Song, L. Gao et.al, "Optimized graph learning using partial tags and multiple features for image and video annotation," IEEE Transactions on Image Processing, vol. 25, no. 11, pp. 4999–5011, 2016.

[15] F. Shen, Y. Xu et.al, "Unsupervised deep hashing with similarity-adaptive and discrete optimization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 12, pp. 3034–3044, 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY