International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Semantic Correlation Based Deep Cross-Modal Hashing For Faster Retrieval

Nikita Bhatt, Amit Ganatra

Abstract*: Due to growth of multi-modal data, large amount of data is being generated. Nearest Neighbor (NN) search is used to retrieve information but it suffers when there is high-dimensional data. However Approximate Nearest Neighbor (ANN) is a searching method which is extensively used by the researchers where data is represented in form of binary code using semantic hashing. Such representation reduces the storage cost and retrieval speed. In addition, deep learning has shown good performance in information retrieval which efficiently handle scalability problem. The multi-modal data has different statistical properties so there is a need to have method which finds semantic correlation between them. In this paper, experiment is performed using correlation methods like CCA, KCCA and DCCA on NMIST dataset. NMIST dataset is multi-view dataset and result proves that DCCA outperforms over CCA and KCCA by learning representations with higher correlations. However, due to flexible requirements of users, cross-modal retrieval plays very important role which works across the modalities. Traditional cross-modal hashing techniques are based on the hand-crafted features. So performance is not satisfactory as feature learning and binary code generation is independent process. In addition, traditional cross-modal hashing techniques fail to bridge the heterogeneous gap over various modalities. So many deep-based cross-modal hashing techniques were proposed which improves the performance in comparison with non-deep cross-modal techniques. Inside the paper, we presented a comprehensive survey of hashing techniques which works across the modalities.

***Index Terms*: Multi-Modal data, Deep CCA (DCCA), cross-modal retrieval, hashing.**

I. INTRODUCTION

Due to advancement of World Wide Web, different types of data like text, images, audio, video are generated which is semantically consistent. Such data is called multi-modal data. As requirements of users are very flexible, need to develop a retrieval system which works across different modalities. Such retrieval is called cross-modal retrieval where users can give any modality as the input [1,3,6,21,27]. In addition, such retrieval provides complementary information which may be useful in decision making or in any recommendation system. Nearest Neighbor (NN) is widely used in information retrieval but very expensive as dimensionality increases. So researchers are focusing on Approximate Nearest Neighbor (ANN) which resolves the problem of NN by giving approximate solution [3, 4]. The indexing scheme called hashing of ANN is widely used which map high-dimensional data to binary code in comparison with tree-based indexing scheme [21, 27]. Such binary representation leads to less time and less space which helps for efficient retrieval [2,4,5]. In order to retrieve information across multi-modal data, multi-modal hashing (MMH) is used. MMH is categorized in two parts: Multi-Source Hashing (MSH) and Cross-Modal Hashing (CMH). But application scenario of MSH is limited in comparison with CMH as all modalities of data might not be present in practical scenario which is prerequisite for MSH. So CMH is used which explore the correlation among modalities to activate the cross-modal similarity search [11,18, 25, 26, 27]. Existing CMH strategies do feature learning and hash code learning as autonomous technique which may not achieve satisfactory result. But as the emerging technique called deep learning has shown promising result in feature generation, it is not only used as a feature extractor but also use as a hash code generator and it is done in single framework [2,3]. Remaining portion of the paper covers different cross-modal retrieval methods. However, it is required to find semantic similarity between multi-modal data for efficient retrieval as they have different statistical properties. There are many methods available in literature which finds semantic similarity between multi-modal data. Here experiment is performed using canonical correlation analysis (CCA), kernel canonical correlation analysis (KCCA) and deep canonical correlation analysis (DCCA).

Manuscript published on 30 August 2019.
***Correspondence Author(s)**
Nikita Bhatt, U & P U Patel Department of Computer Engineering, CSPIT, CHARUSAT, Gujarat, India. E-mail: nikitabhatt.ce@charusat.ac.in
Dr. Amit Ganatra, U & P U Patel Department of Computer Engineering, CSPIT, CHARUSAT, Gujarat, India.

II. STUDY ON CROSS-MODAL RETRIEVAL METHODS

Cross-Modal retrieval system is broadly divided into common subspace-learning and cross-modal hashing methods [31]. In common subspace-learning, different modalities are mapped to the common subspace which preserve the similarity across modalities. For faster retrieval common representation is mapped to binary code using hashing techniques. Remaining portion covers different methods for common subspace-learning and cross-modal hashing [31].

A.Common Subspace-Learning

Submit your manuscript electronically for review. The subspace-learning technique learns a typical subspace in order to preserve the correlations among various modalities where the likeness will be directly calculated [10]. Figure 1 shows how common subspace is generated for various modalities which will be useful for cross-modal retrieval. Once features are extracted for each modality, semantic similarity between different modalities are identified using correlation methods in order to generate common subspace.

When common subspace is generated, it is required to preserve similarity between multi-modal data. Different methods are available which preserves the semantic correlation across the modalities. In CCA, the functions are simple linear maps and weights of the linear map are chosen such a way that it maximizes the correlation. But, CCA fails to discover non-linear correlation between the objects. To overcome this problem, method called kernel CCA (KCCA) was proposed which learns functions from any reproducing kernel Hilbert space (RKHS) and use different kinds of kernels for each view. Compared to linear CCA, KCCA has more complex function space and produces features that could improve performance of linear classifier. But KCCA takes very long time for training and training set must be stored because of nonparametric method and model is more difficult to interpret [8]. To resolve problems, CCA is extended to deep canonical correlation analysis (DCCA) which is a model for data that is multi viewed. The objective of DCCA is to find deep representation mappings such that output mappings are highly correlated [5]. Experiment is performed on NMIST dataset using correlation methods. In order to represent the non-linearity in the data, there is a method called locality preserving CCA (LPCCA) which forces nearby points to be close in the latent space [28]. But these methods do not give good performance in retrieval when within-class variance is very large [9].

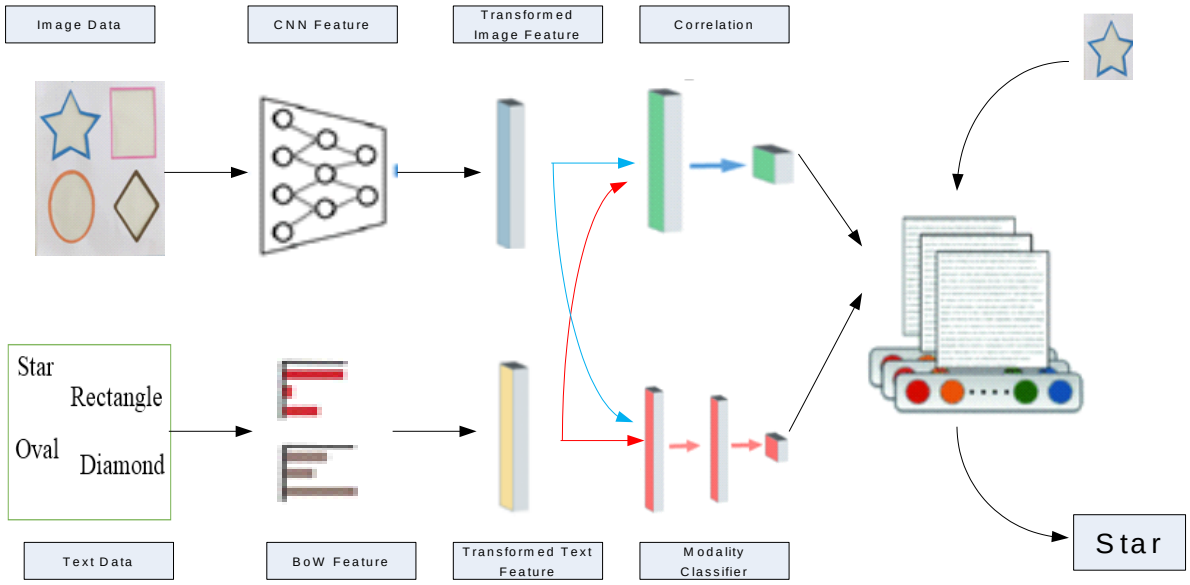


Fig. 1 common subspace-learning

B. Shallow Cross-Modal Hashing

In shallow learning, only single layer of linear and non-linear transformation is performed and to speed up the retrieval process, hashing is widely used. Figure 2 shows cross-modal retrieval using hashing where for each modality binary code is generated using hash function which preserves the similarity between semantically similar modality. In data independent hashing method, learning of the hash code is independent of the training data whereas hash function generation depends on training data in Data-Dependent methods [2,4,5,6]. With Locality sensitive based hashing (LSH) which is data independent method, hash functions are constructed using standard similarity measures like ℓ_p norm, cosine similarity, Jaccard similarity, etc [18]. Although these approaches have asymptotically guaranteed for the very high performance, they can always generate very long binary code. In normal cases, it will generate several hundred binary codes to achieve an acceptable performance.

It is necessary to provide an efficient understanding of multi-modal data considering its fast growing availability with the help of multi-modal hashing. There are many methods used in literature for Cross-Modal retrieval using hashing. In [19], a novel approach was proposed called "Composite hashing with multiple information sources (CHMIS)" which integrates features from multiple sources and apply hashing methods but statistical properties of different sources may be lost. So author has incorporated semantic similarity between hash codes and related hash functions designed for different sources. In [14], author has proposed cross-view hashing (CVH) search where each view is represented as a compact binary codeword. Similar data objects should share the same codeword and retrieval of data objects is possible with codewords having small hamming distance. In addition to have retrieval on muti view data, in [20] proposed an approach which uses multiplication of code-words to discard unwanted embedding and improve accuracy by integrating several information sources of one instance. But it works only when all information sources are available which is not practically possible. Instead of considering single-view, in [21] hash functions are learned by considering multiple-view information.

Retrieval Number: I8157078919/19©BEIESP
DOI: 10.35940/ijitee.I8157.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

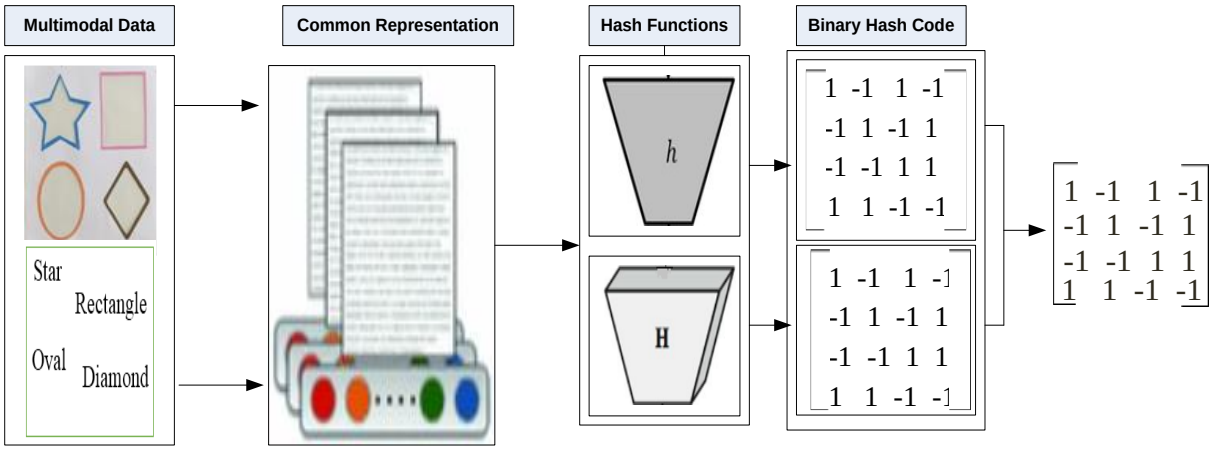


Fig. 2 Cross-Modal Retrieval Using Hashing

In [12], author has proposed a framework for inter-media hashing (IMH) which learns a hash function for each media type and projected into a common hamming space. Inter-Modal consistency is preserved by enforcing the same hash code for relevant media type and Intra-Modal consistency is maintained by conserving local structure details inside every media type. Proposed structure in [13] which is linear cross-modal based hashing (LCMH) divides the training data into k clusters using distance functions which is used to solve scalability issue. To maintain the inter-similarity, data representations is mapped to common binary subspace where binary codes for semantically similar modals are consistent. Traditional methods for cross-view search translate each of the views to one of the views of a multi-view data and apply single-view similarity search. In many practical applications, it is rarely used as translation is application specific and very time consuming.

The limitation of LCMH and CVH is that they heavily rely on Eigen decomposition operations which are very costly when the data dimensionality is very high. Here hash function learning is considered as a generalized eigenvalue problem. In [15], author has proposed multi-modal latent binary embedding (MLBE) in which it is assumed that hash codes are latent factors and use these latent factors along with other model parameters to generate the observations. There are two kinds of observations such as inter-modality similarity (text-image similarity labels) and intra-modality similarity (text-text or image-image similarity values). In [16], a novel method was proposed called co-regularized hashing (CRH) where for each modality; hash function is optimized to reduce the overall loss. As loss function is non-convex, concave convex procedure was applied to resolve the optimization issue. Once the hash function for one bit is learned, CRH learns other bits via boosting procedure such that bias can be minimized. However it will not work fine specifically to large scale data as training time complexity reaches to $O(n^2)$, here n is the size of data points. In [17], proposed method called semantic correlation maximization (SCM) uses supervised data for training time and learn hash function bit by bit so there is no need for hyper-parameters tuning and stopping conditions.

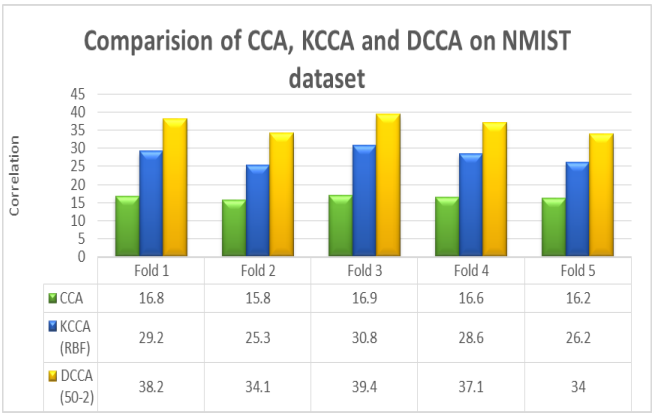
C. Deep Cross-Modal Hashing

Above mentioned techniques including CVH, MLBE, CRH, and CMFH have feature generation as well as hash code generation as two independent processes which might not generate satisfactory results [15, 16, 21]. So deep based cross-modal hashing techniques are currently hot research topic where feature extraction as well as generation of the hash code can be achieved in the common framework. Deep Cross-modal hashing techniques are classified into supervised and unsupervised techniques. In [28], unsupervised cross-modal hashing techniques were proposed where Deep Boltzman Machine was used which maps multi-modal representations into unified representations and model can be useful for classification and retrieval tasks. In [29], author has used stacked auto-encoders for cross-modal retrieval which divides given dataset into mini bunches and different mapping functions for each bunch is adjusted. Another another deep unsupervised method was proposed in [30] which uses multilayer perceptron which creates single representation space across different modalities. But in comparison with unsupervised cross-modal hashing techniques, supervised hashing techniques are more popular due to its performance. The methods for supervised deep cross-modal hashing presented in [22], where compact binary code is generated using deep networks for modalities like image and text. In [3], deep learning based cross-modal hashing (DCMH) framework was proposed which solve discrete optimization problem by directly learning hash codes without relaxation. In addition, here deep networks were used for feature generation along with hash code generation. To reduce really huge semantic gap between the representations of images and texts, [22] proposed a deep based visual semantic hashing (DVSH) approach where as a first step, they used deep learning to get the visual as well as semantic embedding and then fusion operation was performed to get the join hamming embedding for the image-sentence pair. Also, by using pairwise similarity information, the similar image-text pairs may get the similar binary codes.

After getting the fusion codes for each modality, deep learning techniques were used to get the binary code which needs to be almost same as fusion codes and used modality specific model for cross-modal retrieval. Liong, Venice Erin, et al. [23] has proposed a novel network which learns couple of consecutive transformations and learns binary code using cross-modal fusion network. In [24, 25], author has implemented one supervised and one unsupervised algorithms for multi-modal retrieval which learn mapping functions using deep learning techniques which preserve semantic relationships. In [26], author has proposed deep semantic hashing (DSH) method to support cross-modal retrieval which has used fine-tuned CNN model for the images and use fully-connected neural networks for text features extraction. In comparison with the traditional hashing methods which cost higher due to relax discrete constraint, in [27] author has proposed Discrete Latent Semantic Hashing (DLSH) which expanding the efficiency and quantization loss. So in recent years, many deep based cross-modal hashing techniques were proposed for flexible and fast retrieval.

III. Experimental Study and Result Discussion

In multi-modal data, the first challenge is to find semantic correlation among the different types of data. So to find correlation, experiment is performed using CCA, KCCA and DCCA models on NMIST dataset which is multi-view data. We perform experiments with NVIDIA GPU where primary memory is of 64 GB with 4GB of GPU memory. To train a DCCA, we use denoising autoencoder to pertain the layer of each side individually and jointly fine-tune all the parameters to maximize the total correlation. We have two multi-layer perceptron one for each view and at the final output representation is related through CCA. For DCCA, radial basis function (RBF) kernel is used as it is a stationary kernel and hence it is uniform to translation. Whereas, linear kernel doesn't have the stationary property. Apart from that the RBF kernel has one parameter over the polynomial kernel which requires three different parameters. As DCCA 50-2 has the peak performance for k = 50, here in the experiment DCCA uses 50-2 layers where 50 is the output-size and 2 is count of layers used in model. It is observed that a dimension ordering is always preserved in CCA and KCCA but it is not the case with DCCA [7]. To maintain it, here at each target dimensionality new DCCA representation is computed. Figure 3 shows that DCCA consistently finds more correlation at every fold compare to CCA and KCCA on NMIST dataset. For testing, we look at the total correlation on held-out test data and have an output dimensionality of 50 for all CCA, KCCA and DCCA. Fig 3 shows that DCCA has shown consistent correlation captured in 50 most correlated dimensions. Hence DCCA detects more correlation than CCA or KCCA and deeper representations can outperform the shallow ones.



	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
CCA	16.8	15.8	16.9	16.6	16.2
KCCA (RBF)	29.2	25.3	30.8	28.6	26.2
DCCA (50-2)	38.2	34.1	39.4	37.1	34

Fig. 3 Comparison of CCA, KCCA and DCCA on NMIST dataset

IV. Conclusion

Due to large amount of multi-modal data, deep based retrieval techniques are used which can efficiently handle the problem of scalability. These methods use semantic hashing for faster retrieval. Because of flexible requirements received from the users, cross-modal hashing methods are very popular where users may retrieve information of interest by giving any modality as input. These methods map original data points into binary form which leads to less time and space in order to retrieve information. Traditional hashing methods for cross-modal perform feature extraction as well as hash code generation as an independent process but performance is degraded. For the sake of improvement in the performance of retrieval process, deep based cross-modal hashing techniques are used which combines the process of feature extraction and binary code generation in one framework and common subspace is generated where original data points are represented in form of binary code. The common subspace is generated based on various correlation methods which find correlation among various modalities.

REFERENCES

1. K. Wang, Q. Yin, W. Wang, S. Wu and L. Wang, "A Comprehensive Survey on Cross-modal Retrieval", arXiv.org, 2016. [Online]. Available: https://arxiv.org/abs/1607.06215.
2. Q.-Y. Jiang and W.-J. Li, "Deep Cross-Modal Hashing," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
3. Q. Jiang and W. Li, "Asymmetric Deep Supervised Hashing", *arXiv.org*, 2019. [Online]. Available: https://arxiv.org/abs/1707.08325.
4. J. Wang, T. Zhang, N. Sebe and H. Shen, "A survey on learning to hash", *Arxiv.org*, 2018. [Online]. Available: https://arxiv.org/pdf/1606.00185.
5. Q. Li, Z. Sun, R. He and T. Tan, "Deep Supervised Discrete Hashing", *Papers.nips.cc*, 2017. [Online]. Available: https://papers.nips.cc/paper/6842-deep-supervised-discrete-hashing.
6. K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen, "Deep learning of binary hash codes for fast image retrieval," *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015.
7. G. Andrew, R. Arora, J. Bilmes and K. Livescu, "Deep Canonical Correlation Analysis", *PMLR*, 2013. [Online]. Available: http://proceedings.mlr.press/v28/andrew13.html.

8. S. J. Hwang and K. Grauman, "Accounting for the Relative Importance of Objects in Image Retrieval," *Proceedings of the British Machine Vision Conference 2010*, 2010.
9. A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized Multiview Analysis: A discriminative latent space," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
10. T. Sun and S. Chen, "Locality preserving CCA with applications to data visualization and pose estimation," *Image and Vision Computing*, vol. 25, no. 5, pp. 531–543, 2007.
11. R. He, M. Zhang, L. Wang, Y. Ji, and Q. Yin, "Cross-Modal Subspace Learning via Pairwise Constraints," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5543–5556, 2015.
12. J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," *Proceedings of the 2013 international conference on Management of data - SIGMOD 13*, 2013.
13. X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, "Linear cross-modal hashing for efficient multimedia search," *Proceedings of the 21st ACM international conference on Multimedia - MM 13*, 2013.
14. S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search", *Ijcai.org*, 2011. [Online]. Available: https://www.ijcai.org/Proceedings/11/Papers/230.pdf.
15. Y. Zhen and D.-Y. Yeung, "A probabilistic model for multimodal hash function learning," *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 12*, 2012.
16. D.-Y. Y. Yi Zhen, "Co-Regularized Hashing for Multimodal Data," *Co-Regularized Hashing for Multimodal Data*, [Online]. Available: https://papers.nips.cc/paper/4793-co-regularized-hashing-for-multimodal-data.
17. D. Zhang and W. J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," ACM Digital Library. [Online]. Available: https://dl.acm.org/citation.cfm?id=2892854.
18. J. Yu, Y. Lu, Z. Qin, W. Zhang, Y. Liu, J. Tan, and L. Guo, "Modeling Text with Graph Convolutional Network for Cross-Modal Information Retrieval," *Advances in Multimedia Information Processing – PCM 2018 Lecture Notes in Computer Science*, pp. 223–234, 2018.
19. D. Zhang, F. Wang, and L. Si, "Composite hashing with multiple information sources," *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR 11*, 2011.
20. S. Kim, Y. Kang, and S. Choi, "Sequential Spectral Learning to Hash with Multiple Representations," *Computer Vision – ECCV 2012 Lecture Notes in Computer Science*, pp. 538–551, 2012.
21. G. Ding, Y. Guo, and J. Zhou, "Collective Matrix Factorization Hashing for Multimodal Data," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
22. Cao, Yue, Wang, Jianmin, Yang, Qiang, Yuy, and P. S., "Deep visual-semantic hashing for cross-modal retrieval," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, [Online]. Available: http://repository.ust.hk/ir/Record/1783.1-80551.
23. V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Cross-Modal Deep Variational Hashing," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
24. C. Wang, H. Yang, and C. Meinel, "A deep semantic framework for multimodal representation learning," *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 9255–9276, 2016.
25. W. Wang, X. Yang, B. Ooi, D. Zhang and Y. Zhuang, "Effective deep learning-based multi-modal retrieval", *The VLDB Journal*, vol. 25, no. 1, pp. 79-101, 2015. Available: 10.1007/s00778-015-0391-4.
26. Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-Modal Retrieval With CNN Visual Features: A New Baseline," *IEEE Transactions on Cybernetics*, pp. 1–12, 2016.
27. X. Lu, L. Zhu, Z. Cheng, X. Song and H. Zhang, "Efficient discrete latent semantic hashing for scalable cross-modal retrieval", *Signal Processing*, vol. 154, pp. 217-231, 2019. Available: 10.1016/j.sigpro.2018.09.007.
28. N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines", *Jmlr.org*, 2012. [Online]. Available: http://jmlr.org/papers/volume15/srivastava14b/srivastava14b.pdf.
29. W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," *Proceedings of the VLDB Endowment*, vol. 7, no. 8, pp. 649–660, 2014.
30. J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal Similarity-Preserving Hashing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 824–830, 2014.
31. F. Zhong, Z. Chen, G. Min, Z. Ning, H. Zhong, and Y. Hu, "Combinative hypergraph learning in subspace for cross-modal ranking," *Multimedia Tools and Applications*, vol. 77, no. 19, pp. 25959–25982, 2018.

AUTHORS PROFILE

Nikita Bhatt is working at U & P U Patel Department of Computer Engineering in Chandubhai S Patel Institute of Technology, CHARUSAT. She had received degree of Master of Technology in Computer Engineering from Charotar University of Science and Technology and currently pursuing her Ph.D. in the area of Deep Learning. Her research interests include Data Mining, Machine Learning and Deep Learning. She has also published 2 books and more than 15 research papers in the area of data mining and machine learning. She is a member of Board of Studies (BOS) at CHARUSAT. She is also a member of Computer Society of India.

Dr. Amit Ganatra, is working as a Professor at Computer Engineering Department of Charotar University of Science and Technology. He is concurrently holding Deanship in Faculty of Technology-CHARUSAT, Gujarat (since Jan 2011 to till date). He is a member of Board of Studies (BOS), Faculty Board, Academic Council, Research Council and Governing Body for CHARUSAT and member of BOS for CHARUSAT, Gujarat Technological University (GTU), KSV, Indus University, Dr. Bhimrao Ambedkar Open University and C. U. Shah University. He was the founder head of CE and IT departments of CITC (now CSPIT). His research are includes Data Mining, Machine Learning, Artificial Intelligence and Soft Computing. He has more than 50 research publication in reputed journals.