Applied Informatics

RESEARCH Open Access

Cross-media residual correlation learning

Mingkuan Yuan, Xin Huang and Yuxin Peng*

*Correspondence:
pengyuxin@pku.edu.cn
Institute of Computer Science and Technology, Peking University, Beijing, China

Abstract

Due to the progress of deep neural networks (DNN), DNN has been employed to cross-media retrieval. Existing cross-media retrieval methods based on DNN can convert separate representation of each media type to common representation by inter-media and intra-media constraints. By using common representation, we can measure similarities between heterogeneous instances and perform cross-media retrieval. However, it is challenging to optimize common representation learning due to the inter-media and intra-media constraints, which is a multi-objective optimization problem. This paper proposes residual correlation network (RCN) to address this issue. RCN optimizes common representation learning with a residual function, which can fit the optimal mapping from separate representation to common representation and relieve the multi-objective optimization problem. The experiments show that proposed approach achieves the best accuracy compared with 10 state-of-the-art methods on 3 datasets.

Keywords: Cross-media retrieval, Deep residual learning, Representation learning

Introduction

Multimedia retrieval has become an indispensable part of contemporary Internet development. As an important application of artificial intelligence, cross-media retrieval [1] can meet users' requirements to find relevant multimedia data of their queries. For example, a photo of a news event can be submitted by user as a query to retrieve the relevant text materials or video reports. However, considering the inherent discrepancy of instances between different media types, it is challenging to retrieve instances with different media types. The theory of machine learning is introduced to address this issue, which mostly learns common representation of heterogeneous instances and then measures their similarities for cross-media retrieval. These cross-media retrieval methods have mostly verified their effectiveness [2–6].

Canonical correlation analysis (CCA) [2] is a classical method, learning a subspace that maximizes the correlation between different media types. It is extended by Hardoon et al. [3] as kernel canonical correlation analysis (KCCA). Cross-media factor analysis (CFA) [4] is another cross-media retrieval method, which minimizes the Frobenius norm between the pairwise instances in the common space by building the projection functions for different media types. Zhai et al. [7] use metric learning to learn projection functions, and this approach is further improved by adding semi-supervised information

Springer Open

as joint representation learning (JRL) [5]. Kang et al. [6] propose local group based consistent feature learning (LGCFL), which can deal with unpaired data.

Moreover, due to the significant improvement achieved by DNN in representation learning, such as image classification [8], it has been employed to learn the common representations of cross-media instances. Cross-media retrieval methods based on deep neural networks (DNN) have shown their remarkable performance [9–12], making use of DNN's powerful abstraction ability to learn the common representations for different media types. An extension of the restricted Boltzmann machine (RBM) is applied by Ngiam et al. [9] to get shared representation and bimodal autoencoders (Bimodal AE) is proposed, producing common representation for different media types by a shared code layer. Feng et al. [10] propose a method named correspondence autoencoder (Corr-AE) to model the reconstruction and correlation constraints simultaneously. Peng et al. [11] propose the cross-media multiple deep network (CMDN), using hierarchical learning to exploit the complex cross-media correlation.

However, previous methods [5, 9, 10] mostly use inter-media constraints (such as correlation constraints [10]) and intra-media constraints (such as semantic [5] or reconstruction constraints [9, 10]) to build common representations for cross-media retrieval. It is challenging to optimize common representation learning since inter-media and intra-media constraints both need to be considered as objective functions [13, 14], which restrains the performance of cross-media retrieval. He et al. [15] propose deep residual learning and introduce that it is easier to optimize the residual mapping than to optimize the original. If an original mapping is optimal, the residual function could be fitted as zero mapping. If not, it could fit the discrepancy between separate representations and common representations and further optimize the common representation learning. Inspired by this paradigm, we propose residual correlation network (RCN) method to address the above cross-media optimization problem. RCN models the discrepancy between original separate representations and expected common representations by a residual function, which benefits to fitting the optimal mapping by back-propagation procedure at training stage. Then RCN can use the optimal mapping to generate optimal common representations for cross-media retrieval at testing stage.

The rest of this paper is presented as follows "Methods" section introduces the network architecture, objective function, and implementation details of the proposed RCN method. "Experiments" section shows experimental details, results, and analysis. At the end, the conclusion of this paper is concluded in "Conclusions" section.

Methods

To find the optimal solutions of common representation learning under inter-media and intra-media constraints, we use residual learning [15] to model the discrepancy between the original separate representations and expected common representations for optimization. We achieve it by designing the residual correlation network (RCN) based on the convolutional neural networks (CNN) as shown in Fig. 1. It is seen that RCN has two subnetworks: separate representation learning subnetwork and residual correlation learning subnetwork.

In cross-media retrieval problem, we denote the dataset as $D = \{I_L, T_L, I_U, T_U\}$, where $\{I_L, T_L\}$ presents labeled image/text pairs and $\{I_U, T_U\}$ corresponds unlabeled image/

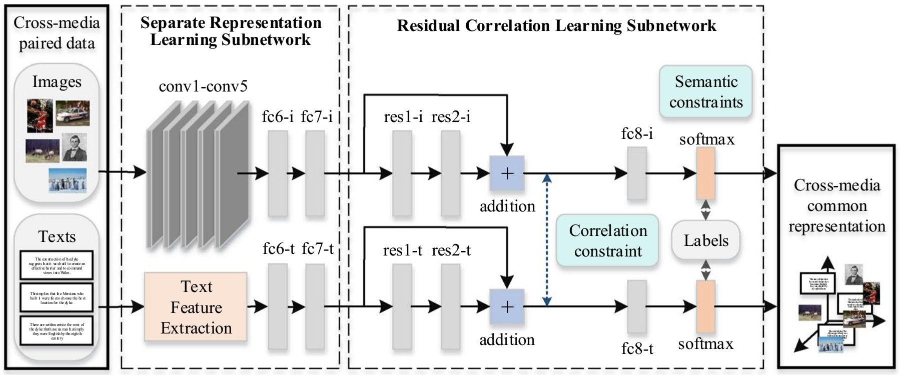

Fig. 1 An overview of our residual correlation network (RCN)

text pairs. The labeled images are denoted as $I_L = \{i_p, y_p\}_{p=1}^{n_1}$, and the labeled texts are denoted as $T_L = \{t_q, y_q\}_{q=1}^{n_1}$. Correspondingly, $I_U = \{i_p\}_{p=1}^{n_2}$ and $T_U = \{t_q\}_{q=1}^{n_2}$ are denoted as the unlabeled images and texts. $\{I_L, T_L\}$ is only used in training stage and $\{I_U, T_U\}$ is only used in testing stage. For residual correlation learning, we denote the separate representations and common representations of an instance x as $f_s(x)$ and $f_c(x)$, which are differed by a residual function $\Delta f(x)$. So the aim of our RCN is to learn common representations of unlabeled images I_U as $R_U^I = \{f_c^I(i_p)\}_{p=1}^{n_2}$ and unlabeled text T_U as $R_U^T = \{f_c^T(t_q)\}_{q=1}^{n_2}$ for cross-media retrieval.

Separate representation learning

There are two pathways in separate representation learning subnetwork: image pathway and text pathway. In the image pathway, we use five convolutional layers (conv1–conv5) and two fully-connected layers (fc6-i and fc7-i) of AlexNet [8] pre-trained on ImageNet [16] from the Caffe [17] Model Zoo. In the text pathway, there are two corresponding fully-connected layers (fc6-t and fc7-t) trained from scratch, which have the same dimension with image fully-connected layers. Moreover, the text pathway receives the BoW feature of each text instance as the original representations. The learning rates of five convolution layers are set at zero to maintain their parameters in training stage, while the learning rates of all the fully-connected layers are set as 0.01.

Residual correlation learning

Existing methods [5, 9, 10] mostly combine inter-media constraints (such as correlation constraints [10]) and intra-media constraints (such as semantic [5] or reconstruction constraints [9]) to train their models for building common representations. Since inter-media and intra-media constraints both need to be optimized as objective functions, there is a complex optimization problem limiting the performance of cross-media retrieval. To address this issue, we propose the residual correlation learning which is illustrated in this section.

In the RCN, as there are inter-media correlation constraints and intra-media semantic constraints converting separate representation $f_s(x)$ of an instance x to common representation $f_c(x)$, $f_s(x)$ and $f_c(x)$ are different, namely $f_c(x) \neq f_s(x)$. So it is reasonable that $f_s(x)$ and $f_c(x)$ can be differed by a residual function $\Delta f(x)$ as:

$$f_c(x) = f_s(x) + \Delta f(x) \tag{1}$$

Instead of designing stacked nonlinear layers to approximate $f_c(x)$ by a correlation constraint as [10], we design several stacked nonlinear layers to approximate the residual function $\Delta f(x) = f_c(x) - f_s(x)$. The process of $f_s(x) + \Delta f(x)$ is realized by a shortcut connection and an element-wise addition, so that the residual function is parameterized by residual layers. By the shortcut connection, the parameters of separate representation learning subnetwork can be directly influenced by the correlation constraint and semantic constraint. If the optimal mapping of common representations can be built by the separate representation learning subnetwork, the representations generated by the residual layers can fit as zero mapping, which makes common representation identical as separate representations. If the subnetwork cannot fit the optimal mapping of common representation, the residual layers can fit the discrepancy between separate representations and common representations and further optimize the common representation learning. Therefore, it is easier to optimize the common representation learning by the residual layers than directly learning the mapping from separate representations to common representations. In other words, the residual layers can adaptively fit the missing part of separate representation learning compared to common representation learning. The residual function $\Delta f(x)$ can be learned with back-propagation algorithm in training stage.

In the implementation, the residual correlation learning subnetwork also consists of two pathways, and each pathway has exactly the same architecture. Here, we take the image pathway as an example to illustrate. There are two fully-connected layers as residual layers (res1-i and res2-i), whose learning rates and dimensions are same as fc6-i and fc7-i. The two residual layers receive the fc7-i separate representations, with which their generated representations are added to generate the common representations. There is a correlation constraint layer between the common representations. After a fully-connected layer (fc8-i) and a softmax layer, we can build a semantic constraint for images' training and the probability vector as common semantic representations for testing.

As for the correlation constraint, our RCN uses Euclidean distance between the representations generated by cross-media residual layers, which can be denoted as:

$$d_c^2(i_p, t_p) = \left\| f_c^I(i_p) - f_c^T(t_p) \right\|^2. \tag{2}$$

Then we get the correlation constraint as:

$$Loss_{\text{Correlation}} = \sum_{p=1}^{n_1} d_c^2(i_p, t_p) = \sum_{p=1}^{n_1} \left\| f_c^I(i_p) - f_c^T(t_p) \right\|^2. \tag{3}$$

Minimizing this correlation constraint can reduce the disparity between representations generated by residual layers to achieve common representation learning. Moreover, RCN also uses the labels of instances to establish the semantic correlation for different media types. The semantic constraints are presented as follows:

$$Loss_{Semantic} = -\frac{1}{m}\sum_{p=1}^{m} f_{\text{sm}}(f_c^I(i_p), y_p, \theta) - \frac{1}{m}\sum_{p=1}^{m} f_{\text{sm}}(f_c^T(t_p), y_p, \theta), \tag{4}$$

where $f_{sm}(i_p, y_p, \theta)$ is the softmax loss function as:

$$f_{\mathrm{sm}}(x, y, \theta) = \sum_{j=1}^{c} 1\{y = j\} log \left(\frac{e^{\theta_j x}}{\sum_{l=1}^{c} e^{\theta_l x}} \right), \tag{5}$$

where θ denotes the network's parameters. The label of instance x is denoted by y and c denotes the number of classes. $1\{y = j\}$ is an indicator function, which equals 1 if $y = j$, otherwise equals 0.

It should be noted that the loss function $Loss_{Semantic}$ is a negative constraint. This is because that if the parameters are closer to the optimization than previous iterations, the prediction results will be closer to the label and the softmax loss function f_{sm} will return a higher positive score. So we need to set the $Loss_{Semantic}$ as a negative constraint.

Therefore, the objective function of RCN can be denoted as:

$$\begin{aligned} \mathrm{Loss}_{\mathrm{Total}} &= \lambda \mathrm{Loss}_{\mathrm{Correlation}} + \mathrm{Loss}_{\mathrm{Semantic}} \\ &= \lambda \sum_{p=1}^{n_1} \left\| f_c^I(i_p) - f_c^T(t_p) \right\|^2 - \frac{1}{m} \sum_{p=1}^{m} f_{\mathrm{sm}}(f_c^I(i_p), y_p, \theta) \\ &\quad - \frac{1}{m} \sum_{p=1}^{m} f_{\mathrm{sm}}(f_c^T(t_p), y_p, \theta), \end{aligned} \tag{6}$$

where λ is a trade-off parameter.

By optimizing with stochastic gradient descent (SGD) algorithm, RCN can minimize this objective function and perform cross-media common representation learning in training stage. As we can see, there is a multi-objective optimization problem in the objective function of RCN. So the residual function can help RCN to converge to a global optimization, which addresses the multi-objective optimization problem. Moreover, due to the fact that the SGD algorithm is stochastic, the iteration complexity is independent of the number of instances [18], which ensures the scalability of RCN.

Experiments

Datasets

The 3 datasets used in the experiment are introduced here. For fairness, we take the dataset splits that are completely same as [10, 11] for our RCN and compared methods.

Wikipedia dataset

Rasiwasia ey al. [19] is the most extensively-used dataset in cross-media retrieval as [5, 10, 11], constructed with the Wikipedia "featured articles". Wikipedia dataset has 2866 image/text pairs and 10 categories: art, biology, geography, history, literature, media, music, royalty, sport, and warfare. The text instance contains several paragraphs that belong to the same section in one Wikipedia web-page with the paired image. The dataset is split randomly into training set of 2173 pairs, test set of 462 pairs, and validation set of 231 pairs.

NUS-WIDE-10k dataset

Feng et al. [10] is a subset of NUS-WIDE dataset [20]. NUS-WIDE dataset contains approximately 270,000 images and each image has their corresponding tags. NUS-WIDE dataset consists of 81 categories, which exist overlaps. NUS-WIDE-10k dataset is composed of 10 largest categories of NUS-WIDE dataset and has 10,000 image/text pairs. The dataset is split randomly into training set of 8000 pairs, test set of 1000 pairs, and validation set of 1000 pairs.

Pascal Sentences dataset

Farhadi et al. [21] contains 1000 images selected from 2008 PASCAL development kit, which belongs to 20 categories. Each image has 5 sentences as the description. Pascal Sentences dataset is split randomly into training set of 800 pairs, test set of 100 pairs, and validation set of 100 pairs.

Compared methods and evaluation settings

There are 10 state-of-the-art compared methods in the experiment: CCA [2], KCCA (with Gaussian kernel and polynomial kernel) [3], CFA [4], JRL [5], LGCFL [6], Multimodal DBN [22], Bimodal AE [9], Corr-AE [10], CMDN [11], and Deep-SM [12]. CCA, KCCA, CFA, JRL, LGCFL are traditional methods without deep learning, and Multimodal DBN, Bimodal AE, Corr-AE, CMDN, and Deep-SM are deep learning methods.

We conduct two retrieval tasks in the experiment: retrieving texts by image query (Image→Text) and retrieving images by text query (Text→Image). Each image in test set is regarded as query individually to retrieve all the text in test set and vice versa. We adopt the MAP score as the evaluation metric, computed for all the compared methods.

As for image input settings, RCN is an end-to-end processing network and receives the raw images directly as input. Nevertheless, except Deep-SM and RCN, all compared methods could only receive extracted feature as input. For fairness, we use the fc7 layer feature extracted from fine-tuned AlexNet whose architecture is same as our RCN. For text, we adopt BoW feature that is completely same as [10, 11] for our RCN and all compared methods. The dimension of BoW feature for Wikipedia dataset is 3000, and the dimensions of BoW feature for Pascal sentences and NUS-WIDE-10k datasets are both 1000.

As for validation set, Multimodal DBN, Bimodal AE, Corr-AE, and CMDN require it for parameter optimization while our RCN also needs it for determining the number of iterations, but the other methods don't. So in the training and test stages, validation set is not used except Multimodal DBN, Bimodal AE, Corr-AE, CMDN, and our RCN.

Experimental results

The MAP scores of 10 compared methods and our RCN are shown in Table 1. Above all, our RCN achieves the best MAP score on all 3 datasets. Compared with the state-of-the-art method Deep-SM, our RCN accomplishes an exciting enhancement from 0.402 to 0.453 on Wikipedia dataset. RCN achieves the highest MAP score of 0.507 on NUS-WIDE-10k dataset. RCN's result on Pascal Sentences dataset is also the best, compared with the other methods.

Table 1 MAP scores of our RCN and compared methods

Dataset	Method	Task		
		Image→Text	Text→Image	Average
Wikipedia dataset	CCA	0.176	0.178	0.177
	CFA	0.330	0.306	0.318
	KCCA (Poly)	0.230	0.224	0.227
	KCCA (Gaussian)	0.357	0.328	0.343
	Bimodal AE	0.301	0.267	0.284
	Multimodal DBN	0.204	0.145	0.175
	Corr-AE	0.373	0.357	0.365
	JRL	0.408	0.353	0.381
	LGCFL	0.416	0.360	0.388
	CMDN	0.409	0.364	0.387
	Deep-SM	0.458	0.345	0.402
	RCN (OnlyCorrelation)	0.465	0.407	0.436
	Our RCN	0.489	0.418	0.454
NUS-WIDE-10k dataset	CCA	0.159	0.189	0.174
	CFA	0.299	0.301	0.300
	KCCA (Poly)	0.129	0.157	0.143
	KCCA (Gaussian)	0.295	0.162	0.229
	Bimodal AE	0.234	0.376	0.305
	Multimodal DBN	0.178	0.144	0.161
	Corr-AE	0.306	0.340	0.323
	JRL	0.410	0.444	0.427
	LGCFL	0.408	0.374	0.391
	CMDN	0.410	0.450	0.430
	Deep-SM	0.389	0.496	0.443
	RCN (OnlyCorrelation)	0.360	0.406	0.383
	Our RCN	0.497	0.517	0.507
Pascal Sentences dataset	CCA	0.110	0.116	0.113
	CFA	0.341	0.308	0.325
	KCCA (Poly)	0.271	0.280	0.276
	KCCA (Gaussian)	0.312	0.329	0.321
	Bimodal AE	0.404	0.447	0.426
	Multimodal DBN	0.438	0.363	0.401
	Corr-AE	0.411	0.475	0.443
	JRL	0.416	0.377	0.397
	LGCFL	0.381	0.435	0.408
	CMDN	0.458	0.444	0.451
	Deep-SM	0.440	0.414	0.427
	RCN (OnlyCorrelation)	0.433	0.443	0.438
	Our RCN	0.472	0.453	0.463

Moreover, we conduct a baseline experiment named “RCN (OnlyCorrelation)”, which means that we design the correlation network without residual layers. Comparing RCN (OnlyCorrelation) and our RCN, it can be seen that the residual correlation learning can actually optimize the common representation learning and lead to much better performance of cross-media retrieval, especially on the NUS-WIDE-10k dataset.

As for the stability of compared methods, we can see that their performances have different trends in the 3 datasets. For instance, the result of Deep-SM is higher than JRL on Wikipedia dataset, while the trend is different on Pascal Sentences dataset. In contrast, our RCN carries on the highest MAP scores on the all 3 datasets, showing the generality of residual correlation learning method, and it effectively learns better common representation and enhances the cross-media retrieval accuracy.

Conclusions

This paper proposes a new cross-media retrieval method RCN, enhancing the optimization of common representation learning for cross-media retrieval. RCN models the discrepancy between separate representations and common representations by a residual function and further optimizes the common representation learning. Moreover, adequate compared experiments are conducted on 3 extensively-used cross-media datasets, which verify the promising capability of our approach.

In the future work, we will extend our RCN to more complex optimization situations which have more objective functions in order to make full use of its promising optimization ability. Moreover, we intend to verify the performance of our RCN under unsupervised setting, further improving its generalization in more applications.

Authors' contributions

The authors discussed the problem and proposed the methods all together. All authors participated in drafting and revising the final manuscript. All authors read and approved the final manuscript.

Acknowledgements
None.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Not applicable.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding
This work was supported by National Natural Science Foundation of China under Grants 61771025 and 61371128.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 31 August 2017   Accepted: 25 September 2017
Published online: 05 October 2017

References

1. Peng Y, Huang X, Zhao Y (2017) An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. In: IEEE transactions on circuits and systems for video technology. IEEE, New Jersey
2. Hotelling H (1936) Relations between two sets of variates. Biometrika 28(3/4):321–377
3. Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: an overview with application to learning methods. Neural comput 16(12):2639–2664
4. Li D, Dimitrova N, Li M, Sethi IK (2003) Multimedia content processing through cross-modal association. In: Proceedings of the eleventh ACM international conference on multimedia. p 604–611
5. Zhai X, Peng Y, Xiao J (2014) Learning cross-media joint representation with sparse and semisupervised regularization. IEEE Trans Circ Syst Video Technol 24(6):965–978

6. Kang C, Xiang S, Liao S, Xu C, Pan C (2015) Learning consistent feature representation for cross-modal multimedia retrieval. IEEE Trans Multimed 17(3):370–381
7. Zhai X, Peng Y, Xiao J (2013) Heterogeneous metric learning with joint graph regularization for cross-media retrieval. AAAI, California
8. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. p 1097–1105
9. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011) Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11). p 689–696
10. Feng F, Wang X, Li R (2014) Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM International Conference on Multimedia, New York. p 7–16
11. Peng Y, Huang X, Qi J (2016) Cross-media shared representation by hierarchical learning with multiple deep networks. IJCAI, New York, pp 3846–3853
12. Wei Y, Zhao Y, Lu C, Wei S, Liu L, Zhu Z, Yan S (2017) Cross-modal retrieval with cnn visual features: a new baseline. IEEE Trans Cybern 47(2):449–460
13. Coello CAC, Lamont GB, Van Veldhuizen DA (2007) Evolutionary algorithms for solving multi-objective problems, vol 5. Springer, New York
14. Marler RT, Arora JS (2004) Survey of multi-objective optimization methods for engineering. Struct Multidiscip Optim 26(6):369–395
15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, New Jersey. p 770–778
16. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition, CVPR 2009, New Jersey. p 248–255
17. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia. p 675–678
18. Sra S, Nowozin S, Wright SJ (2012) Optimization for machine learning. Mit Press, Cambridge
19. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GR, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM international conference on multimedia. p 251–260
20. Chua T-S, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. p 48
21. Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D (2010) very picture tells a story: gnerating sentences from images. In: European Conference on Computer Vision, Springer. p 15–29
22. Srivastava N, Salakhutdinov R (2012) Learning representations for multimodal data with deep belief nets. In: International conference on machine learning workshop