7CCEMPRJ

Robot Cross-Sensory Retrieval using Deep Neural Networks

Wenjie Zhang

K21113522

k21113522@kcl.ac.uk

Supervisor: Dr.Shan Luo

King’s College London

Department of Engineering

February 16, 2023

CONTENTS

LIST OF FIGURES

***Abstract*—With the development of artificial intelligence, researchers have started to pay more attention to developing general-purpose artificial intelligence. In order to fulfil this purpose, developing multi-sensory object-centric perception, reasoning and interaction is inevitable. However, research on these parts is limited compared to other artificial intelligence areas, such as computer vision. Most previous related work focus on images-to-text or text-to-images, this project mainly explores ‘cross-sensory(vision, tactile and audio) retrieval using deep neural network’ by the Stanford ObjectFolder dataset.**

***Index Terms*—ObjectFolder, cross-sensory retrieval, deep neural network**

I. Introduction

“Modality” is a biological concept introduced by German physiologist Hermann von Helmholtz. It is the channel through which creatures receive information by virtue of their perceptual organs and experiences, e.g. humans have a visual modality, an auditory modality, etc.

A. Background, Motivations and Problem statement

In contemporary data-driven artificial intelligence research, the information provided by a single modality is no longer sufficient to enhance the cognitive capabilities of machines. Similar to humans who utilise multi-sensory information, for example, vision, auditory sense and touch, to perceive the world, machines’ cognitive ability could be facilitated by multi-sensory perception as well. Cross-modal retrieval, also known as cross-media retrieval, is characterised by the presence of data from all modalities of the training set, but only one modality is available during the testing process. The aim is to achieve information faculty interaction between modalities and mining the relationships between samples from different modalities, retrieving samples from one modality with approximate semantics from another modality as a query. For instance, the first thing to start the morning is turning off the alarm clock. As reaching for an alarm clock, we already know what shape we are expecting, even if our eyes are not open yet, which is the audio-to-tactile retrieval process that is already done in our brain. The physical properties, such as 3d shape, impact sound and material types, of different objects lead to unique sensory data. In 1976, MCGURK H, MACDONALD H. Hearing lips and seeingvoices[J]. Nature, 1976, 264(5588): 746-748. proposed the influence of vision on speech perception, which was later used in audio-visual speech recognition (AVSR) called the prototype multi-modal concept. In 2010, fusion methods and fusion levels as clues to classify existing multi-modal fusion research methods In 2015, Multi-modal hide-condition random fields are proposed to improve the classification of multi-modal data; deep multi-modal hashing methods based on orthogonal regularisation constraints are proposed to reduce the information redundancy of multi-modal representations. In 2019, Delineating multi-modal learning research directions: multi-modal representation, multi-modal translation, multi-modal alignment, multi-modal fusion, multi-modal co-awareness, etc.

Most previous research mainly works on image-to-text retrieval or text-to-images retrieval, research related to vision, audio and tactile sensory data retrieval is rather limited. The main problem that this project needs to tackle is how to extract features from each modality into a common latent space and then retrieve the specific sensory data by querying from the latent space produced. There are some side problems that need to be solved, the most obvious, the sparsity of feature space due to the high dimensionality.

B. Aims and Objectives

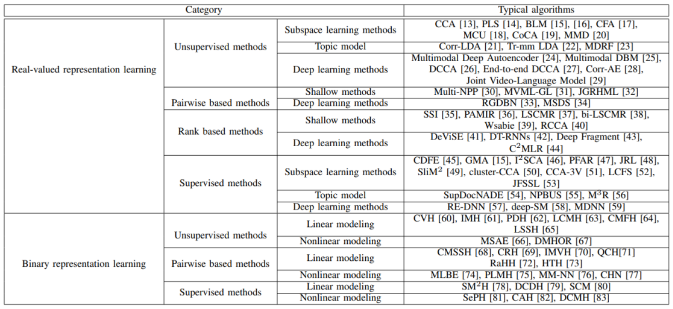
The aim of this project is to implement a baseline method as illustrated in the [1] along with a newly developed approach to perform cross-sensory retrieval on the Stanford ObjectFolder dataset and evaluate the effectiveness of approaches.

C. Proposed methodologies and expected results

There are two major ways to implement cross-sensory retrieval, 1. Real-valued representation learning and 2. Binary representation learning. Both can be further divided into supervised/unsupervised methods, pairwise-based methods and rank-based approaches. This project mainly focuses on an unsupervised real-valued representation learning approach, in particular, the Subspace learning method which measures the similarities between different sensory data. We would expect the distance between different modalities of the same/similar object to be small and vice versa. Thus, cross-sensory data can be retrieved.

II. Literature review

There are plenty of approaches in the field of cross-sensory retrieval, as introduced above, these approaches can be categorized into two 1. Real-valued representation learning and 2. Binary representation learning. This part will briefly introduce several methods in each category.



Category			Typical algorithms
Real-valued representation learning	Unsupervised methods	Subspace learning methods	CCA [13], PLS [14], BLM [15], [16], CFA [17], MCU [18], CoCA [19], MMD [20]
		Topic model	Corr-LDA [21], Tr-mm LDA [22], MDRF [23]
		Deep learning methods	Multimodal Deep Autoencoder [24], Multimodal DBM [25], DCCA [26], End-to-end DCCA [27], Corr-AE [28], Joint Video-Language Model [29]
	Pairwise based methods	Shallow methods	Multi-NPP [30], MVML-GL [31], JGRHML [32]
		Deep learning methods	RGDBN [33], MSDS [34]
	Rank based methods	Shallow methods	SSI [35], PAMIR [36], LSCMR [37], bi-LSCMR [38], Wsabie [39], RCCA [40]
		Deep learning methods	DeViSE [41], DT-RNNs [42], Deep Fragment [43], C^2MLR [44]
	Supervised methods	Subspace learning methods	CDFE [45], GMA [15], I^2SCA [46], PFAR [47], JRL [48], SliM^2 [49], cluster-CCA [50], CCA-3V [51], LCFS [52], JFSSL [53]
		Topic model	SupDocNADE [54], NPBUS [55], M^{3}R [56]
		Deep learning methods	RE-DNN [57], deep-SM [58], MDNN [59]
Binary representation learning	Unsupervised methods	Linear modeling	CVH [60], IMH [61], PDH [62], LCMH [63], CMFH [64], LSSH [65]
		Nonlinear modeling	MSAE [66], DMHOR [67]
	Pairwise based methods	Linear modeling	CMSSH [68], CRH [69], IMVH [70], QCH[71] RaHH [72], HTH [73]
		Nonlinear modeling	MLBE [74], PLMH [75], MM-NN [76], CHN [77]
	Supervised methods	Linear modeling	SM^{2}H [78], DCDH [79], SCM [80]
		Nonlinear modeling	SePH [81], CAH [82], DCMH [83]

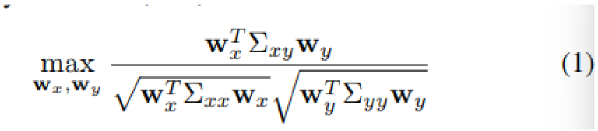
Fig. 1. list of methods.

A. Real-valued representation learning

An object has different sensory representations, for instance, a cup can be represented by impact sound, texture or a picture. What is expected in common is these different modality data of an object that share the same feature representation space tend to have a short distance from each other. Real-valued representation learning means that during training a representation space can be produced. The different sensory data can be measured by real values in the space. There are some subdivision methods categories:

1) Unsupervised: The common representations are learnt by co-occurrence information in the multi-modal data, for example, multi-sensory data appearing in the same object folder tends to represent the same object. This can be further divided into subspace learning technical and deep learning approaches.

For subspace learning methods, how to measure the similarity of cross-sensory data is a key issue. The objective of this method is to produce a latent space that contains different modality data which then can be calculated to produce their similarities. The most famous technical of subspace learning is Canonical correlation analysis (known as CCA) which, in short, constructs relationships between cross-sensory data. Fig.2 shows the correlation calculation formula where xx/yy just means the covariance matrix of two sensory data.



$$\max_{\mathbf{w}_x,\mathbf{w}_y} \frac{\mathbf{w}_x^T \Sigma_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \Sigma_{xx} \mathbf{w}_x}\sqrt{\mathbf{w}_y^T \Sigma_{yy} \mathbf{w}_y}} \qquad (1)$$

Fig. 2. equation.

There are some other popular technical in cross-sensory retrieval as well, such as 1. partial least squares (PLS) and bilinear model (BLM). A highly correlated common linear space is produced from different modality data by Sharma and Jacobs utilizing PLS. In addition, a latent space for cross-modal face recognition is produced by Tenenbaum and Freeman using the bilinear model. Furthermore, A maximum covariance unfolding (MCU) method is proposed by Mahadevan et al., which reduces the dimensionality of different modal data. Collective component analysis (CoCA) is proposed by Shi et al. to reduce the heterogeneous latent space dimensionality. The sparse projection matrices method is proposed by Wang et al. to do cross-modal retrieval, which transforms image-text data into a latent space.

As deep learning becomes a hot topic, Ngiam et al. are inspired who train deep neural networks on multi-sensory data (mainly speech audio coupled with video) to produce a representation space. After that, multi-modal joint representations are produced by deep Restrict Boltzmann Machine, which learns separate modal low-level representations and then merges together to produce a joint representation. Most recently, a model involving a correspondence autoencoder is proposed by Feng et al. for cross-sensory retrieval. The Corr-AE combines hidden representations of two uni-modal autoencoders. The losses are correlation learning errors and representation learning errors, which need to be minimised.

2) Supervised: Supervised methods have one more advantage compared to unsupervised methods, which is they exploit label information. Label information could better separate objects in the common representation space.

For the subspace learning method, fig shows the difference between supervised learning and unsupervised learning. There are many different kinds of CCA technical, such as cluster-CCA, three-view CCA, and multi-label CCA. Despite CCA-based approaches, a common discriminant feature extraction (CDFE) is proposed by Lin and Tang to obtain the common feature subspace. The difference within and between the scatter matrix is maximized. Moreover, a feature learning approach is proposed by Zhai et al to explore the information of correlation in an optimization framework for cross-sensory data, named Joint Representation learning (JRL).

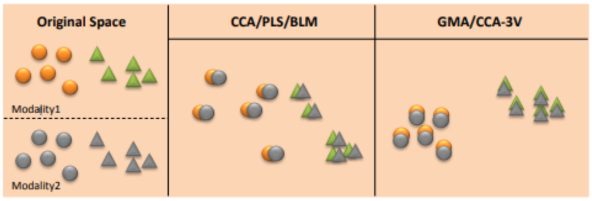


Fig. 3. comparison.

For deep learning methods, a regularized deep neural network (RE-DNN) is proposed by Wang et al. to map semantics in different modalities. The neural network contains five layers to map multi-modal features into a latent space. So that the similarity can be calculated between different sensory data. After that, a deep convolutional neural network model is proposed by Wang et al. combined with a neural language model to map images and text, named multi-modal deep neural network (MDNN) which is robust against noise.

III. Methodologies

There are plenty of approaches available which can be the baseline of this project. In this project, CCA is chosen to be the baseline as same as the Stanford ObjectFolder project, as we are using the ObjectFolder dataset. It will be more feasible to evaluate this project.

A. Dataset

As artificial intelligence gets more and more attention, a huge number of data has been generated. In the multi-sensory retrieval field, there are numerous datasets, such as Wikipedia, INRIA-Websearch, Flickr30K, Pascal VOC and NUS-WIDE datasets. However, they are not suitable for this project, as either, they only contain a small number of objects or only contain image and text data. For instance, Flickr30K only contains 31783 images along with sentences written by natives, with no audio or tactile data. Wiki, INRIA-Websearch and NUS-WIDE are the same. Thus, the ObjectFolder dataset is chosen for this project, which contains 1000 objects along with impact sound, tactile reading and appearance. ObjectFolder datasets encode the 1000 objects by implicit neural representation network—ObjectFile in form of implicit neural representations. Each representation encodes three sensory data for that object, visual, acoustic and tactile.

B. *Baseline and model*

As mentioned above, Canonical correlation analysis is implemented as a project baseline to evaluate effectiveness. At the end of CCA, it learns two modal directions x and y, in which the correlation between x and y is maximised. The original CCA has some disadvantages so there are numerous variants of CCA, as shown in Fig4. For example, the signs of canonical correlations are indeterminate and it requires N more than Pk. However, the advantages of CCA are the reason why it is chosen as a baseline approach, such as easy to apply, invariant to scaling, etc.

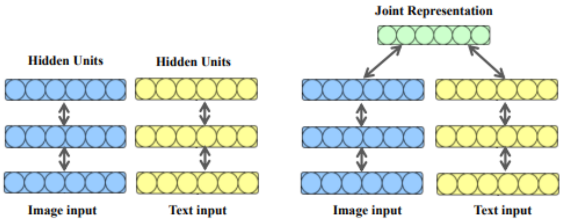


Fig. 4. structure.

The deep canonical correlation analysis model proposed by Yan and Mikolajczyk[1], as shown in fig, might be a good start to implement our own model. The objective function is Fig5.

$$\max_{\mathbf{W}_x,\mathbf{W}_y} tr(\mathbf{W}_x^T \Sigma_{xy} \mathbf{W}_y)$$
$$s.t. : \mathbf{W}_x^T \Sigma_{xx} \mathbf{W}_x = \mathbf{W}_y^T \Sigma_{yy} \mathbf{W}_y = I$$

Fig. 5. equation.

IV. Project schedule

I've tried to include a Gantt diagram, but Latex seems not to allow it. Therefore, I just put a simplified work table here. This workflow leaves two months in case any part needs more time.

TASK	ASSIGNED TO	PROGRESS	START	END
Phase 1 Title				
writing abstract	Name	0%	16/2/23	23/2/23
writing intro		0%	23/2/23	5/3/23
gathering/reading papers		0%	5/3/23	15/3/23
writing literature review		0%	15/3/23	29/3/23
implement baseline methods			30/3/23	27/4/23
Phase 2 Title				
developing new approach		0%	27/4/23	4/5/23
implement and testing new approach		0%	4/5/23	1/6/23
experiment			1/6/23	6/6/23
writing methodologies			6/6/23	13/6/23
writing experiment			13/6/23	20/6/23
Phase 3 Title				
writing conclusion and future work			20/6/23	25/6/23

V. Risk, feasibility and ethic

As this project mainly works from home by using a computer, it should have no ethical issues or risks. (Apart from getting Covid/flu accidentally on the underground) This project should be feasible if the time allocated is reasonable.

References

[1] Kaiye Wang, Qiyue Yin†, Wei Wang, Shu Wu, Liang Wang, Senior Member, IEEE, "A Comprehensive Survey on Cross-modal Retrieval". 2016

[2] Guanqun Cao, Yi Zhou, Danushka Bollegala and Shan Luo , Spatio-temporal Attention Model for Tactile Texture Recognition, 2020.

[3] Zhenghao Liu1 Chenyan Xiong2 Yuanhuiyi Lv1 Zhiyuan Liu3 Ge Yu1, "UNIVERSAL VISION-LANGUAGE DENSE RETRIEVAL: LEARNING A UNIFIED REPRESENTATION SPACE FOR MULTI-MODAL RETRIEVAL," 2023, pp. 271–350.

[4] Ruohan Gao Yen-Yu Chang Shivani Mall Li Fei-Fei Jiajun Wu, "OBJECTFOLDER: A Dataset of Objects with Implicit Visual, Auditory, and Tactile Representations " 2021.

[5] Ruohan Gao1 Zilin Si2 Yen-Yu Chang1 Samuel Clarke1 Jeannette Bohg1 Li Fei-Fei1 Wenzhen Yuan2 Jiajun Wu1, "OBJECTFOLDER 2.0: A Multisensory Object Dataset for Sim2Real Transfer" 2022.