Multimodal Data and Causal Representation Learning.

Yi-Jing Sie

Abstract

Multimodal data has become increasingly popular in recent years because in the real life situation, information entailed by a phenomenon can be acquired by varied instruments and stored in different forms of modality. Moreover, it is rare to see that a single modality contains comprehensive information of the phenomenon of interest. The goal of this report is to provide a brief overview of the fundamental concepts as well as some high-level ideas in multimodal data fusion and multi-task learning as a means to introduce and motivate the needs of learning causal representation for multimodal data.

1 Motivation

In the real world situation, the information that underlies data is normally stored/collected in different modalities. For example, videos are usually associated with voices and captions, where both the voices and the captions carry part of the information about the subject of interest that is not captured by the videos. Furthermore, different statistical features distinguish different modalities. Images, for example, are typically represented as pixel intensities, but texts are typically represented as discrete word count vectors. Therefore, due to these distinct statistical features of different information resources, it is critical to discover the interactions and relationships among different modalities, that is the underlying causal structures.

2 Introduction

Multimodal data refers to the feature of data such that it can collected by a variety of ways and represented by different measurable parameters. For example, image, video, speech, text, and so on. Although merging different modalities or forms of information to improve model performance appears to be intuitively appealing, it is difficult to combine the varying levels of noise and conflicts between modalities in practice. So far, there has been no definite solutions to handle these difficulties, however research has been focusing on utilizing data augmentation, pre-training models, self-supervised learning, and so on.

However, I would like to argue that such strategies may not be sufficient, and that it is necessary to learn the underlying causal model rather than only statistical connections, such as correlation, between variables. I believe this is the case because by knowing the true causal model for the given multimodal data, we allow distribution shifts which is one of the common challenges in multimodal data.

3 Related Work

3.1 Overview of Multimodal Data Fusion

This part is mainly an overview of this paper[9], and will briefly answer why multimodal data fusion is needed, how to perform it, and some challenges at the model design level and the data level. *Audiovisual multimodality* refers to the fusion of audio and vision, two of our most informative senses. For example, human speech and communication both heavily rely on the information stored in these two senses because our brains integrate speech information represented by the waveforms and visual information about what is being said represented by the movements of lips to process the message information. Such interaction between hearing and vision in speech perception can be demonstrated by the McGurt effect [10], which emphasizes the importance of integrating both visual and audio information to enhance the overall model performance on tasks such as speech recognition.[11][6]

So now we know that in real life scenarios, information usually comes from multiple sources, which can be divided into sources of interest that carry valuable information and other sources which do not carry any information of interest. The latter one is often considered as noise when we are processing the data. In the context of strategies for multimodal data fusion, there are two types of methods that can be considered: one is model driven method, and the other is data driven method. Model driven approaches rely on explicit and realistic description of models, so it is generally successful when the assumed model and assumptions hold.[1][4] Nonetheless, model driven methods might not be ideal when the assumptions are too complicated or the models are unknown.As a consequence, data driven approaches are preferred because it prompts us to make the fewest assumptions and use the simplest models when performing multimodal data fusion, where simple model assumptions can mean, for example, linear relationships between variables.

To incorporate diversity of multimodal data in the models, a concrete mathematical formula can be written as follows:

$$x_{ij} = \sum_{r}^{R} a_{ir} b_{jr} \tag{1}$$

An common interpretation is that at sample index $j = 1, 2, 3, \cdots, J$, x_{ij} is a linear combination of R signals $b_{j1}, b_{j2}, \cdots, b_{jR}$ collected via sensor $i = 1, 2, 3, \cdots, I$ with weights $a_{i1}, a_{i2}, \cdots, a_{iR}$. The corresponding matrix form is

$$X = AB^T \tag{2}$$

where $X \in \mathbb{K}^{I \times J}$, $\mathbb{K} = \{\mathbb{R}, \mathbb{C}\}$, and $A \in \mathbb{K}^{I \times R}$, $B \in \mathbb{K}^{J \times R}$. Unfortunately, we often cannot retrieve the underlying factor matrices A and B because for any $R \times R$ invertible matrix T, the following formula always holds:

$$X = AB^T = (AT^{-1})(TB^T) \tag{3}$$

This implies that the pair $\{(AT^{-1}), (TB^T)\}$ have the same contribution to the observations X and thus cannot be distinguished from the true underlying factor matrix-pairs $\{A, B\}$. One of the most popular approaches is to consider T as a unitary matrix by making an assumption that the columns of B are decorrelated. In such cases, the indeterminacy in formula(3) can be reduced to the rotation problem [3][8][7]. To tackle the indeterminacy in (3), one of the most well-known approaches is independent component analysis (ICA), which is more commonly formulated as:

$$X = AS \tag{4}$$

where X is composed of the observed random vectors, A is known as full rank mixing matrix , and S is composed of statistically independent sources. Details can be found in [13]. As we can

infer from the above formula, if all data sets exist a (multi)linear relationship and share the same underlying factorization model, it is possible to use a single matrix or tensor decomposition for data fusion.

Though the acquisition of multimodal data is not a problem in real life thanks to the advance technology development, there are still some complex issues and challenges in the processing of multimodal data. For example, data collected in real world settings suffered from a variety of uncertainty, and the presence of heterogeneous multiple datasets entails new types of uncertainty.

Regarding challenges at the model design level, When we want to use models for prediction learning from multimodal data, there are two key difficulties to overcome:

- Iintra-modal and Cross-modal interactions has to be learnt in order to make reasonable prediction
- Models must be robust to unexpected missing or noisy modalities during testing.

On the other hand, challenges at the data level are involved in the process of pre-processing multimodal data for models. For example, in practice, not all observations or datasets have the same level of confidence, reliability or information quality [14], [2], [9], which emphasizes the importance of balancing information from different sources.

3.2 Multi-task Learning

This part mainly references this paper [12].Multi-task learning (MTL) is a subfield of machine learning where multiple tasks are solved at the same time by a shared model exploiting commonalities and differences across tasks. This enables the generalization of our model by using the domain information contained in the training datasets. Since multi-task learning using a shared representation to simultaneously learn multiple tasks, it is suitable to be incorporated for learning causal representation for multimodal data.

In terms of deep learning models, multi-task learning can be done with either hard or soft parameter sharing in the models' hidden layers.

- **Hard Parameter Sharing** refers to the strategy of sharing the hidden layers for all tasks, while retaining some task-specific output layers.

  As we can see from the Figure 1, parameters in the hidden layers of the model is shared for learning all tasks; however, the output layers are separated for task-specific outputs. Hard parameter sharing is effective in preventing overfitting because the shared hidden layers learn to capture all the common information across different tasks with shared parametric representation, while separated task-specific output layers distinguish the output representation from different tasks. Therefore, the more tasks the model learns, the less the chance is for model to overfit.

- **Soft Parameter Sharing**, on the other hand, keeps different models for different tasks, but the difference between the parameters from different models is constrained by making use of regularization to force similar parameters among different models.A variety of regularization strategies can be chosen for this task. For example, [5] uses L2 norm, while [15] employs trace norm.

  As we can see from the Figure 2, parameters in the hidden layers of the model are separated while learning across tasks; however, the hidden layers from different models are connected to enforce similarity parametric representation across different tasks.

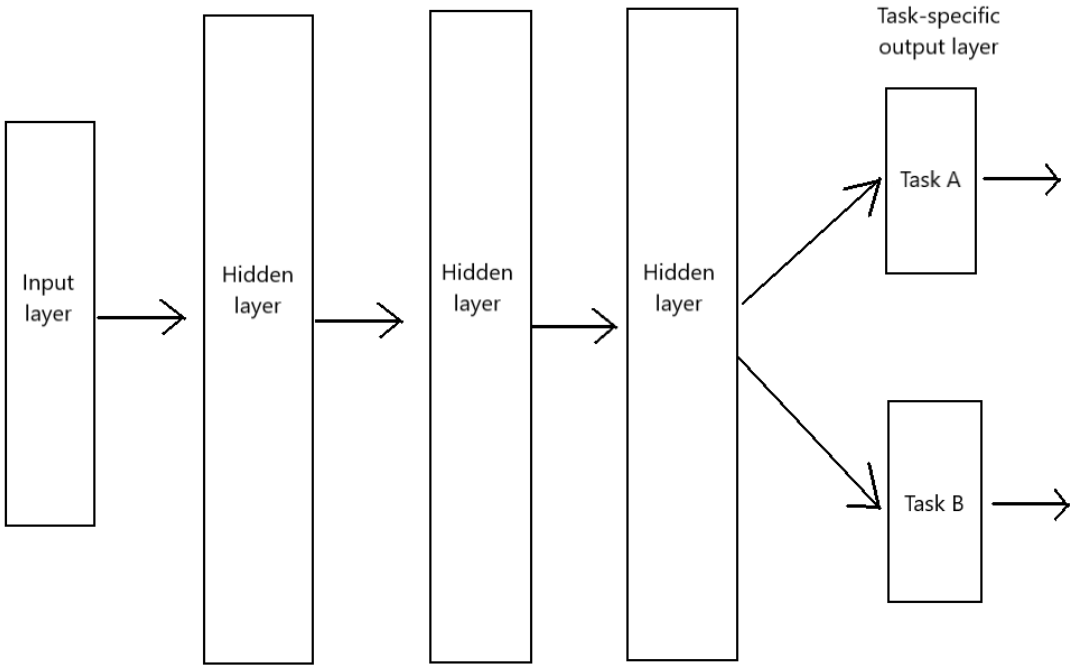

Figure 1: Hard parameter sharing in deep neural networks

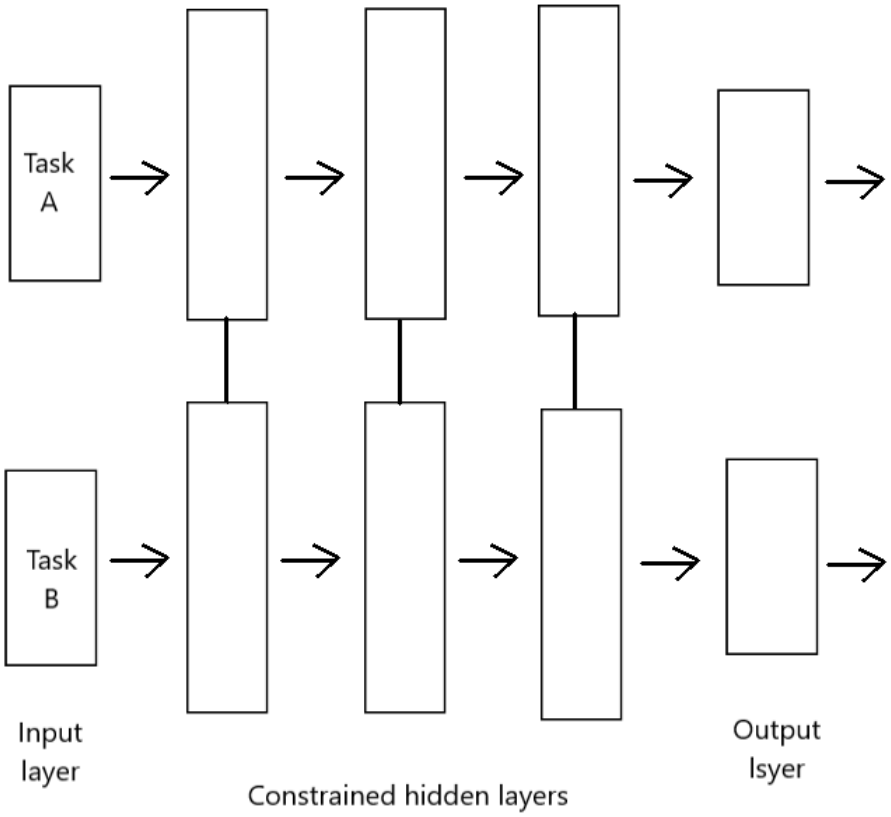

Figure 2: Soft parameter sharing in deep neural networks

One of the goals in multi-task learning is to enable models to learn the shared representation across different tasks of interest. There are different goals in multi-task learning as well, but for the purpose of learning causal representation for multimodal data, which is the main objective of this report, I am not going into details about them.

4 Conclusion

As more and more advanced technology developed, the acquisition of multimodal data becomes much easier than before. However, the challenges of processing, utilizing, and fusing the information stored in different modalities are still perplexing nowadays researchers and scientists. In this report, I briefly reviewed some basic concepts, methods and challenges associated with multimodal data fusion as well as the fundamental ideas of multi-task learning as an introduction of learning causal representation of multimodal data. If we can find a way to discover the underlying causal structure from the provided multimodal datasets, we are closer to enter a new era where deep learning and machine learning models have higher generalizability and hence more powerful and applicable than ever.

References

[1] Felix Biessmann et al. "Analysis of multimodal neuroimaging data". In: *IEEE reviews in biomedical engineering* 4 (2011), pp. 26–58.

[2] Vince D Calhoun and Tulay Adali. "Feature-based fusion of medical imaging data". In: *IEEE Transactions on Information Technology in Biomedicine* 13.5 (2008), pp. 711–720.

[3] Raymond B Cattell. ""Parallel proportional profiles" and other principles for determining the choice of factors by rotation". In: *Psychometrika* 9.4 (1944), pp. 267–283.

[4] Nicolle M Correa et al. "Canonical correlation analysis for data fusion and group inferences". In: *IEEE signal processing magazine* 27.4 (2010), pp. 39–50.

[5] Long Duong et al. "Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser". In: *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)*. 2015, pp. 845–850.

[6] Dafydd Gibbon, Inge Mertins, and Roger K Moore. "Audio-visual and multimodal speech-based systems". In: *Handbook of Multimodal and Spoken Dialogue Systems*. Springer, 2000, pp. 102–203.

[7] Richard A Harshman et al. "Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multimodal factor analysis". In: (1970).

[8] Joseph B Kruskal. "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics". In: *Linear algebra and its applications* 18.2 (1977), pp. 95–138.

[9] Dana Lahat, Tülay Adali, and Christian Jutten. "Multimodal data fusion: an overview of methods, challenges, and prospects". In: *Proceedings of the IEEE* 103.9 (2015), pp. 1449–1477.

[10] Harry McGurk and John MacDonald. "Hearing lips and seeing voices". In: *Nature* 264.5588 (1976), pp. 746–748.

[11] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. "Deep multimodal learning for audio-visual speech recognition". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, pp. 2130–2134.

[12] Sebastian Ruder. "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (2017).

[13] James V Stone. "Independent component analysis: an introduction". In: *Trends in cognitive sciences* 6.2 (2002), pp. 59–64.

[14] Tom F Wilderjans, Eva Ceulemans, and Iven Van Mechelen. "The SIMCLAS model: Simultaneous analysis of coupled binary data matrices with noise heterogeneity between and within data blocks". In: *Psychometrika* 77.4 (2012), pp. 724–740.

[15] Yongxin Yang and Timothy M Hospedales. "Trace norm regularised deep multi-task learning". In: *arXiv preprint arXiv:1606.04038* (2016).