Quantum-inspired Multimodal Representation

Qiuchi Li qiuchili@dei.unipd.it Department of Information Engineering University of Padova Padova, Italy

ABSTRACT

We introduce our work in progress that targets on building multimodal representation under quantum inspiration. The challenge for multimodal representation falls on a fusion strategy to capture the interaction between different modalities of data. As the most successful approaches, neural networks lack a mechanism of explicitly showing how different modalities are related to each other. We address this issue by seeking inspirations from Quantum Theory (QT), which has been demonstrated advantageous in explicitly capturing the correlations between textual features. In this paper, we give an overview of the related works and present the proposed methodology.

KEYWORDS

multimodal data fusion, quantum physics, neural networks

1 INTRODUCTION

In human communication, messages are often conveyed through a combination of different modalities, such as visual, audio and linguistic modalities. In order for an automatic understanding of the multimodal messages, one needs to fuse the information from different modalities to construct a joint multimodal representation. The challenge falls on how to characterise the interactions between different modalities, which can be complicated in some scenarios. In the example shown in Fig. 1, the visual-linguistic query cannot be understood solely by either the text or image individually, but the relation between the image and text must be correctly recognized. This is where the significance and challenge sit for multimodal representation learning.

Most existing research focus on neural network-based fusion strategies for constructing multimodal representation, ranging from the earlier Hidden Markov Model (HMM)-based models [7] to different RNN variants [1] and more recently tensor-based approaches [12] and seq-to-seq structures [8]. Despite their strong accuracy performances, the interplay between different data modalities are encoded in an inherent way by the neural network components, making it difficult for humans to understand the contributions of each modality in a particular task.

Quantum Theory (QT) provides a well-established theoretical and mathematical formalism for describing the physical world on a microscopic level. Beyond physics, QT frameworks have been applied to many research areas [2, 9, 10, 15], among which successful Massimo Melucci melo@dei.unipd.it Department of Information Engineering University of Padova Padova, Italy

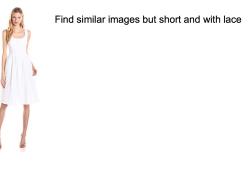


Figure 1: Example of a Multi-modal query.

results have been observed in text-based language understanding [10, 15]. In this context, quantum-inspired frameworks have the natural advantage of explicitly capturing the correlations between features by the concepts of *quantum superposition* and *quantum entanglement*.

This inspires us to adopt quantum-inspired frameworks for constructing multimodal representation, and propose effective and interpretable multimodal fusion methods. In particular, we propose to capture the interactions within a single modality with *quantum superposition*, and model the cross-modal interactions by means of *quantum entanglement*. In this way, we establish a pipeline that extracts the interactions within multi-modal data in a way understandable from a quantum perspective. We expect to obtain comparable performances to state-of-the-art systems in concrete multimodal tasks.

2 PROPOSED METHODOLOGY

Here we introduce our proposed approach for building multimodal data representation inspired by quantum theory. Essentially, we represent multimodal data as many-body systems composed of different data modalities as subsystems. The interaction of different modalities is inherently captured by the notion of entanglement between the subsystems. We propose to build a complex-valued learning network to implement the quantum theoretical framework, which facilitates learning the cross-modal interactions in a datadriven fashion.

2.1 Complex-valued Unimodal Representation

Complex values are essential for the mathematical formalism of quantum physics. However, most existing quantum-inspired models for text representation are based on the real vector space, ignoring the complex-valued nature of quantum notions. Recently, our prior works [3, 4, 11] leveraged quantum *superposition* to model

Copyright @ 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). *IIR 2019, September 16–18, 2019, Padova, Italy*

IIR 2019, September 16-18, 2019, Padova, Italy

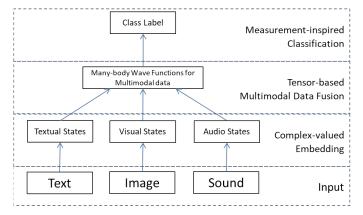


Figure 2: Our proposed multimodal representation framework.

correlations between textual features, and the complex-valued representation leads to improved performance and enhanced interpretability. We attempt to employ the concept of *superposition* for modeling intra-modal interactions, and investigate the complexvalued embedding approach to capture the interactions within other modalities.

2.2 Tensor-based Approaches for Capturing Inter-modal Interactions

We represent multimodal data as a many-body quantum system in entanglement. The mathematical formulation will be a complexvalued tensor constructed from unimodal complex-valued vectors by tensor-based approaches. Tensor-based models have been applied for classification [5] and matching [14] tasks, but they avoid directly computing the tensor by decomposing the tensor and learning the decomposed weights via neural networks. Explicit tensor combination of different data modalities into a holistic multimodal representation remains unexplored. [6] proposed a framework to investigate the entanglement between user and document for relevance feedback, but it remains a challenge on how to apply it to multimodal data.

2.3 Quantum-inspired Framework for Multimodal Sentiment Analysis

We focus on the multimodal sentiment analysis task, and work with the benchmarking datasets CMU-MOSI [13] and CMU-MOSEI [1]. The task is to classify sentiment of a video into 2, 5 or 7 classes with textual, visual and acoustic features. As is shown in Fig. 2, our framework represents unimodal data as a set of pure states through complex-value embedding, and constructs the many-body state of an video utterance through tensor-based approaches. Finally, quantum-like measurement operators are implemented for sentiment classification. The whole process is implemented into a complex-valued neural network, and the parameters in the pipeline can be learned from labeled data in an end-to-end manner.

The network is born with an advantage in interpretability compared to classical neural networks, in that the role of each component is made explicit prior to the network training phase. We are working on deploying the network to CMU-MOSI and CMU-MOSEI, and we expect to see comparable values to state-of-the-art models in terms of effectiveness.

ACKNOWLEDGMENTS

This PhD project is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 721321.

REFERENCES

- [1] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [2] Jerome R. Busemeyer and Peter D. Bruza. Quantum Models of Cognition and Decision. Cambridge University Press, New York, NY, USA, 1st edition, 2012.
- [3] Qiuchi Li, Sagar Uprety, Benyou Wang, and Dawei Song. Quantum-Inspired Complex Word Embedding. Proceedings of The Third Workshop on Representation Learning for NLP, pages 50–57, 2018.
- [4] Qiuchi Li, Benyou Wang, and Massimo Melucci. CNM: An Interpretable Complexvalued Network for Matching. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4139–4148, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [5] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2247–2256, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [6] Massimo Melucci. Towards modeling implicit feedback with quantum entanglement. Quantum Interaction 2008, 2008.
- [7] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. In Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI '11, pages 169–176, New York, NY, USA, 2011. ACM.
- [8] Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabas Poczos. Seq2seq2sentiment: Multimodal Sequence to Sequence Models for Sentiment Analysis. arXiv:1807.03915 [cs, stat], July 2018. arXiv: 1807.03915.
- [9] C. J. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, New York, NY, USA, 2004.
- [10] Alessandro Sordoni, Jian-Yun Nie, and Yoshua Bengio. Modeling Term Dependencies with Quantum Language Models for IR. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, pages 653–662, New York, NY, USA, 2013. ACM.
- [11] Benyou Wang, Qiuchi Li, Massimo Melucci, and Dawei Song. Semantic Hilbert Space for Text Representation Learning. In *The World Wide Web Conference*, WWW '19, pages 3293–3299, New York, NY, USA, 2019. ACM. event-place: San Francisco, CA, USA.
- [12] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. arXiv:1707.07250 [cs], July 2017. arXiv: 1707.07250.
- [13] Amir Zadeh, Rown Zellers, and Eli Pincus. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. page 10.
- [14] Peng Zhang, Zhan Su, Lipeng Zhang, Benyou Wang, and Dawei Song. A Quantum Many-body Wave Function Inspired Language Modeling Approach. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, pages 1303–1312, New York, NY, USA, 2018. ACM.
- [15] Yazhou Zhang, Dawei Song, Peng Zhang, Panpan Wang, Jingfei Li, Xiang Li, and Benyou Wang. A quantum-inspired multimodal sentiment analysis framework. *Theoretical Computer Science*, April 2018.