

UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

**Common and Unique Feature Learning for
Data Fusion**

by

Sunny Verma

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Sydney, Australia

2020

Certificate of Authorship/Originality

I, Sunny Verma declare that this thesis, is submitted in fulfilment of the requirements for award of Doctoral of Philosophy, in the School of Software, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in this thesis.

This document has not been submitted for qualification at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date: 02/03/2020

Sunny Verma

Acknowledgements

First and foremost, I want to thank my supervisors Dr. Wei Liu and Dr. Chen Wang, for their support throughout my degree. The journey as a Ph.D. student is not for everyone. Only you know about the feeling of joy after acceptance of your work and the feeling of lost after a rejection. I could not have achieved half of what I have without strong supervision of my supervisors.

I would like to thank Prof. Liming Zhu and Data61, CSIRO for providing me the scholarship for this degree. I'm also thankful to my former supervisor Dr. Meng Hui Lim for the opportunity under his guidance in Hong Kong, and to many things beyond words.

I also appreciate the support and discussion with my friends, in particular, Dr. Abhinav Anand, Dr. Rajan Kashyap, Dr. Avinash Singh, and Yurui Ming. In a similar spirit, my colleagues at UTS and Data61, CSIRO have been a constant source of valuable moments I can remember.

Finally, I owe my family the irreversible time and patience they have had during all these years and thank Sau Kei Kwok for all her support.

Sunny Verma
Sydney, Australia, 2020.

List of Publications

Journal Papers

- J-1. **Sunny Verma**, Chen Wang, Liming Zhu, and Wei Liu, “Attn-HybridNet: Enhancing Discriminability of Features with Attention Fusion,” *IEEE Transactions on Cybernetics*, **Under Review**.

Conference Papers

- C-1. **Sunny Verma**, Chen Wang, Liming Zhu, and Wei Liu, “DeepCU: Integrating Both Common and Unique Latent Information for Multimodal Sentiment Analysis, *IJCAI19*, Aug 2019. **Accepted**.
- C-2. **Sunny Verma**, Chen Wang, Liming Zhu, and Wei Liu, “Towards Effective Data Augmentation via Unbiased GAN Utilization, *PRICAI19*, Aug 2019. **Accepted**.
- C-3. **Quan Do**, **Sunny Verma**, Fang Chen, and Wei Liu, “Multiple Knowledge Transfer for Cross Domain Recommendation, *PRICAI19*, Aug 2019. **Accepted**.
- C-4. **Sunny Verma**, Chen Wang, Liming Zhu, and Wei Liu, “HybridNet: Improving Deep Learning Networks via Integrating Two Views of Images, *ICONIP18*, Dec 2018. **Accepted**. **Accepted**.

Conference Posters

- P-1. **Sunny Verma**, Chen Wang, Liming Zhu, and Wei Liu, “A Compliance Checking Framework for DNN Models, *IJCAI19*, Aug 2019. **Accepted**.
- P-2. **Sunny Verma**, Chen Wang, Liming Zhu, and Wei Liu, “Extracting Highly Effective Features for Supervised Learning via Simultaneous Tensor Factorization, *AAAI17*, Feb 2017. **Accepted**.

Contents

| | |
|---|----------|
| Certificate | ii |
| Acknowledgments | iii |
| List of Publications | iv |
| List of Figures | ix |
| List of Tables | xi |
| Abstract | xiii |
| 1 Introduction | 1 |
| 1.1 Outline | 2 |
| 1.2 Contributions of this Thesis | 4 |
| 1.3 Thesis Organisation | 6 |
| 2 Background | 7 |
| 2.1 Multimodal Data Fusion | 7 |
| 2.1.1 Early Fusion | 8 |
| 2.1.2 Late Fusion | 9 |
| 2.2 Tensor | 10 |
| 2.2.1 Tensor Preliminaries | 11 |
| 2.2.2 Tensor Decomposition Algorithms | 13 |
| 2.3 Generative Adversarial Networks | 16 |
| 3 Attn-HybridNet: Enhancing Discriminability of Hybrid | |

| | |
|--|-----------|
| Features with Attention Fusion | 18 |
| 3.1 Introduction | 18 |
| 3.2 Our Contributions | 19 |
| 3.3 Literature Review | 20 |
| 3.3.1 Background | 22 |
| 3.3.2 Tensor Factorization using <i>LoMOI</i> | 25 |
| 3.4 The Tensor Factorization Network | 26 |
| 3.4.1 The First Layer | 27 |
| 3.4.2 The Second Layer | 28 |
| 3.5 The Hybrid Network | 29 |
| 3.5.1 The First Layer | 31 |
| 3.5.2 The Second Layer | 32 |
| 3.6 Proposed attention-based fusion Attn-HybridNet | 35 |
| 3.6.1 Computational Complexity | 37 |
| 3.7 Experiments and Results | 37 |
| 3.7.1 Experimental Setup | 37 |
| 3.7.2 Datasets | 38 |
| 3.8 Results and Discussions | 40 |
| 3.9 Summary | 49 |
| | |
| 4 DeepCU: Integrating both Common and Unique Latent Information for Multimodal Sentiment Analysis | 51 |
| 4.1 Introduction | 51 |
| 4.2 Our Contributions | 54 |
| 4.3 Related Work | 55 |

| | | |
|-------|--|----|
| 4.3.1 | Tensor Fusion Networks (TFN) | 56 |
| 4.3.2 | Low-rank Multimodal Fusion (LMF) | 56 |
| 4.3.3 | Hybrid - DeepShallow | 57 |
| 4.4 | Proposed Methodology | 57 |
| 4.4.1 | Unique Network | 59 |
| 4.4.2 | Common Network | 61 |
| 4.4.3 | Fusion Layer | 64 |
| 4.4.4 | Complexity Analysis | 65 |
| 4.5 | Experimental Settings | 65 |
| 4.5.1 | Dataset | 65 |
| 4.5.2 | Baselines | 66 |
| 4.5.3 | Parameter Settings in DeepCU | 68 |
| 4.5.4 | Evaluation Metrics | 69 |
| 4.5.5 | Results and Explainability Analysis | 69 |
| 4.5.6 | Case Study with Missing Values from the Acoustic Modality in the CMU-MOSI Dataset | 73 |
| 4.6 | Summary | 75 |

5 Towards Effective Data Augmentations via Unbiased GAN Utilization **76**

| | | |
|-------|---|----|
| 5.1 | Introduction | 76 |
| 5.2 | Our Contributions | 79 |
| 5.3 | Realted Work | 79 |
| 5.3.1 | GAN utilization in dataset augmentation and their limitations | 81 |
| 5.3.2 | Dataset Bias and GANs | 82 |

| | | |
|----------|---|------------|
| 5.4 | Data Augmentation Pursuit | 83 |
| 5.4.1 | Stage-1. | 84 |
| 5.4.2 | Stage-2. | 85 |
| 5.5 | Experiments | 86 |
| 5.5.1 | Experimental Setup and Performance Metric | 86 |
| 5.5.2 | Feature Extraction for ensemble classifier | 88 |
| 5.6 | Results and Discussions | 88 |
| 5.6.1 | How does data-augmentation affect the performance of classifier? | 89 |
| 5.6.2 | How does the percentage of input data affect the quality of data-augmentation? | 92 |
| 5.7 | Summary | 99 |
| 6 | Conclusion and Future Work | 100 |
| 6.1 | Conclusions of the Thesis | 100 |
| 6.2 | Recommendations for Future Work | 102 |
| | Bibliography | 104 |

List of Figures

| | | |
|-----|---|----|
| 2.1 | Workflow of Early Fusion Scheme. | 9 |
| 2.2 | Workflow of Late Fusion Scheme. | 10 |
| 2.3 | A third order tensor, \mathfrak{X} | 11 |
| 2.4 | Fibers of third order tensor \mathfrak{X} | 12 |
| 2.5 | Slices of third order tensor \mathfrak{X} | 12 |
| 2.6 | Workflow of Generative Adversarial Networks. | 16 |
| 3.1 | Comparison of convolution outputs from Layer1 in PCANet and TFNet on CIFAR-10 dataset. These plots demonstrate the contrast between the two types of information obtained with the two views of the data. | 30 |
| 3.2 | Workflow of the Attn-HybridNet model. | 31 |
| 3.3 | Comparison of factorization strength in Layer 2 of the PCANet , TFNet and HybridNet on CIFAR-10 dataset | 33 |
| 3.4 | Performance Comparison by varying size of the training data | 41 |
| 3.5 | Accuracy of Attn-HybridNet on CIFAR-10 dataset by varying the dimension of attention context vector w in Alg. 3. | 44 |
| 3.6 | (Best viewed in color) Accuracy of various methods on CIFAR-10 dataset by varying size of the training data | 47 |

| | | |
|-----|---|----|
| 3.7 | (Best viewed in color) t-SNE visualization of features from HybridNet (top) and Attn-HybridNet (bottom) on CIFAR-10 dataset. | 48 |
| 4.1 | A typical Multimodal Sentiment Analysis System | 52 |
| 4.2 | Comparison of missing values (interrogation mark ‘?’) scenarios by State of the art A. Low-rank Multimodal Fusion (LMF), B. Tensor Fusion Networks (TFN), and C. our Proposed DeepCU. | 53 |
| 4.3 | Performance comparison of DeepCU vs common (Com) and unique (Unq) networks on the CMU-MOSI dataset. | 70 |
| 4.4 | Performance of DeepCU, TFN, and LMF by varying hyperparameters on the CMU-MOSI dataset. The legend DeepCU-x-y represents, x = number of convolution filters and y = filter size. | 71 |
| 5.1 | Images from PPB dataset [11]. | 77 |
| 5.2 | Procedure of Generating Sieved Synthetic Data | 83 |
| 5.3 | Accuracy-plot of top 6 pairs from Table 5.1. x -axis in subplots represents the values of α used in experiments, where $x = 0, 1$ corresponds to <i>Org</i> , <i>DCGAN</i> . The y -axis represents the mean accuracy obtained after 3-fold <i>crossvalidation</i> | 93 |
| 5.6 | Accuracy-plot of top 6 pairs from Table 5.1. x -axis in subplots represents the values of α used in experiments, where $x = 0, 1$ corresponds to <i>Org</i> , <i>DCGAN</i> . The y -axis represents the mean accuracy obtained after 3-fold <i>crossvalidation</i> | 96 |

List of Tables

| | | |
|-----|--|----|
| 3.1 | Comparison of different feature extraction models | 23 |
| 3.2 | Classification Error (%) obtained by varying hyper-parameters on CIFAR-10 dataset without augmentation | 42 |
| 3.3 | Classification Error (%) obtained on MNIST variations datasets . . . | 45 |
| 3.4 | Classification Error (%) obtained on CURET datasets | 46 |
| 3.5 | Classification Error (%) obtained on CIFAR-10 dataset without data augmentation | 49 |
| 4.1 | Comparison of multimodal data fusion models | 58 |
| 4.2 | Performance comparison of DeepCU vs other fusion techniques on CMU-MOSI dataset. The mean and variance for each baseline and DeepCU are obtained by executing them for five times. This superiority of DeepCU is specifically visible in the case of 7-class classification. | 72 |
| 4.3 | Performance comparison on the POM dataset. | 73 |
| 4.4 | Affect of missing values on DeepCU _{DF} , TFN, and LMF. These feature vectors are taken from the actual CMU-MOSI dataset. | 74 |

- 5.1 Performance comparisons using CNN classifier on baselines datasets and augmented dataset obtained using **DAP**. The *p-values* obtained using *t-tests* on pairs ‘Baseline vs **DAP**’ are tabulated in the last column. Note that we follow the scientific notation ** where we use $1Ex$ to present 1×10^x . Please note that, for the bias and the variance lower value is better whereas for accuracy higher is better. . 90
- 5.2 Performance comparison using SVM classifier on baselines datasets and augmented dataset obtained using **DAP**. The *p-values* obtained using *t-tests* on pairs ‘Baseline vs **DAP**’ are tabulated in the last column. Note that we follow the scientific notation ** where we use $1Ex$ to present 1×10^x . Please note that, for the bias and the variance lower value is better whereas for accuracy higher is better. . 91

ABSTRACT

Common and Unique Feature Learning for Data Fusion

In today's era of big data, information about a phenomenon of interest is available from multiple acquisitions. Data captured from each of these acquisition frameworks are commonly known as modality, where each modality provides information in a complementary manner. Despite the evident benefits and plethora of works on data fusion, two challenging issues persist, 1) feature representation: how to exploit the data diversity that multiple modalities offer, and 2) feature fusion: how to combine the heterogeneous information for better decision making.

To address these challenges, this thesis presents a significantly improved model of two widely utilised fusion techniques, a) early fusion: combining features from multiple modalities for joint prediction, and b) late fusion: combining modality-specific predictions at the decision level. I illustrate how both these techniques have their own specific limitations, with late fusion unable to harness the inter-modality benefits, and the reliance of early fusion on a single model causing failure when information from any modality is futile. To overcome these drawbacks, I developed novel multimodal systems that performs feature extraction and feature fusion in a consolidated frameworks. Technically, I designed feature extraction schemes to capture both unique information from individual modalities and common information from multimode representations. I then combine these two kinds of information for supervised prediction, by designing efficient fusion schemes that enable this frameworks to perform information discovery and feature fusion simultaneously.

In this thesis, I also demonstrated the benefits of fusing both the common and unique information in supervised learning and validate the significance of the developed techniques on multimodal, multiview, and multisource datasets. The designed methods leverage the multimodal benefits by creating additional diversity, and ob-

tain a more unified view of the underlying phenomenon for better decision making.