Multi modal representation learning and cross-modal semantic matching

Wang, X.

Citation

Version: Publisher's Version

License: Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden

Downloaded from: https://hdl.handle.net/1887/3391031

Note: To cite this publication please use the final published version (if applicable).
MULTI MODAL REPRESENTATION LEARNING
AND CROSS-MODAL SEMANTIC MATCHING

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir.H.Bijl,
volgens besluit van het college voor promoties
te verdedigen op vrijdag 24 juni 2022
klokke 11.15 uur

door

Xue WANG
geboren te Heilongjiang, China
in 1989
Promotor:
Prof. Dr. Ir. F. J. Verbeek

Co-promotors:
Dr. Y. Du  (Xi’an Jiaotong University)
Dr. S. Verberne

Promotiecommissie:
Prof. Dr. A. Plaat
Prof. Dr. N. Mentens
Prof. Dr. M.S. Lew
Prof. Dr. H. Trautman  (University of Munster)
Dr. Y. Guo  (Chinese Academy of Sciences)

Copyright © 2022 Xue Wang

The research is financially supported by the Chinese Scholarship Council (CSC No.201906280464).
## CONTENTS

1 **Introduction**
   - 1.1 Datasets for Multi-modality studies
   - 1.2 Importance of Cross-modal Supervision for Representation Learning
   - 1.3 Introduction of Common Semantic Space Representation Learning
   - 1.4 Research Questions and Perspectives
   - 1.5 Thesis Structure

2 **Embedded Representation of Relation Words with Visual Supervision**
   - 2.1 Our Model
     - 2.1.1 Basic CBOW Model
     - 2.1.2 Visually Supervised Word2Vec Model
   - 2.2 Experimental Results
     - 2.2.1 Dataset and Experiment Settings
     - 2.2.2 Results and Analysis
   - 2.3 Conclusion

3 **Fine-Grained Label Learning in Object Detection with Weak Supervision of Captions**
   - 3.1 Related Work
     - 3.1.1 Lexico-semantic Analysis
     - 3.1.2 Weakly Supervised Multiple Instance Learning
     - 3.1.3 Curriculum Learning
   - 3.2 Methodology
     - 3.2.1 Overview
     - 3.2.2 Semantic Mapping
     - 3.2.3 Fine-grained Label Learning Based on Curriculum Learning
   - 3.3 Experimental Results and Discussion
     - 3.3.1 Experimental Setup
     - 3.3.2 Evaluation Metrics
     - 3.3.3 Performance and Analysis
   - 3.4 Conclusion and Future Work

4 **Kernel-Based Mixture Mapping for Image and Text Association**
   - 4.1 Related Work
   - 4.2 Linear Models and the Ineffectiveness
4.3 Proposed Model ................................................. 55
  4.3.1 Local Linear Mapping .................................... 55
  4.3.2 Kernel-based Mixture Mapping ......................... 56
  4.3.3 Constraints in the Model ............................... 58
4.4 Optimization, Algorithm and Analysis ...................... 60
  4.4.1 Optimization and Algorithm .......................... 60
  4.4.2 Convergence Analysis ................................. 62
  4.4.3 Complexity Analysis ................................... 64
4.5 Experimental Results ........................................ 64
  4.5.1 Dataset and Experimental Setting ..................... 64
  4.5.2 Parameter Tuning and Analysis ....................... 67
  4.5.3 Performance on Cross-media Retrieval ................. 70
4.6 Conclusions .................................................. 76

5 Visual Representation Contextualization Based on Contrastive Learning .................................. 79
5.1 Related Work .................................................. 82
  5.1.1 Phrase Grounding ....................................... 82
  5.1.2 Non-maximum Suppression (NMS) ....................... 83
  5.1.3 Contrastive Learning in Cross-modal Tasks .......... 83
5.2 Methodology .................................................. 84
  5.2.1 Overview ................................................. 84
  5.2.2 Visual Representation Contextualization Model .... 85
  5.2.3 Mixed Contrastive Loss Function ...................... 87
5.3 Experimental Results ......................................... 88
  5.3.1 Datasets and Metrics ................................... 88
  5.3.2 Implementation Details ............................... 89
  5.3.3 Quantitative Results .................................. 90
  5.3.4 Ablation Study ......................................... 91
  5.3.5 Qualitative Results ................................... 93
5.4 Conclusion .................................................... 95

6 Conclusions and Discussion .................................... 97
6.1 Main Contributions .......................................... 98
6.2 Achievements of Research Presented in This Thesis ...... 101
6.3 Future Research .............................................. 101

Summary ......................................................... 119
Samenvatting .................................................... 121
Curriculum Vitae ............................................... 123
Acknowledgements ............................................. 125