Pengzhi Gao

CONTACT INFORMATION

Baidu, Inc.
10 Xibeiwang East Road,
Haidian District, Beijing, China 100193

Mobile: (+86) 15810512592
Email: gpengzhi@gmail.com
Homepage: https://gpengzhi.github.io/

EDUCATION

Rensselaer Polytechnic Institute, Troy, NY, USA

Ph.D., Electrical Engineering, August 2013 - December 2017

- Advisor: Professor Meng Wang
- Thesis: High-dimensional Data Analysis by Exploiting Low-dimensional Models with Applications in Synchrophasor Data Analysis in Power Systems

University of Pennsylvania, Philadelphia, PA, USA

M.S., Electrical Engineering, August 2011 - May 2013

Xidian University, Xi'an, China

B.S. (with honors), Electronic and Information Engineering, August 2007 - May 2011

RESEARCH INTERESTS

My research interests lie in the intersection of the fields of signal processing, high-dimensional statistics, natural language processing, and machine learning, with a particular emphasis on developing low-dimensional models, optimization methods, and training strategies with applications in power system analysis, image processing, and machine translation.

WORK EXPERIENCE

Staff Research & Development Engineer September 2020 to Present

Natural Language Processing Department,
Baidu, Inc., Beijing, China
Supervisor: Dr. Zhongjun He

- Proposed a simple but effective training strategy, CrossConST, to boost the zero-shot performance of multilingual neural machine translation (ACL 2023 Findings). Developed training pipeline based on CrossConST and improved the multilingual machine translation system that supports more than 200 languages in Baidu Translate.
- Proposed a simple but effective training strategy, Bi-SimCut, to boost the performance of neural machine translation (NAACL 2022). Developed training pipeline based on Bi-SimCut and improved the machine translation system in Baidu Translate.
- Proposed a novel method, MixDiversity, to generate different translations with high faithfulness and diversity (EMNLP 2021 Findings). Developed data augmentation method based on MixDiversity and improved the machine translation system in Baidu Translate.

Data Scientist / Machine Learning Engineer February 2018 to April 2020

Machine Learning Team,
Petuum, Inc., Pittsburgh, PA, USA
Supervisor: Dr. Tong Wen and Dr. Zhiting Hu

- Designed and implemented the machine learning library (based on TensorFlow, DyNet, and LightGBM) for Petuum AI Builder Platform.
- Designed and developed Texar-PyTorch (https://github.com/asyml/texar-pytorch, gaining over 730 stars), an open-source machine learning and text generation toolkit based on PyTorch.
- Maintained and contributed to Texar (https://github.com/asyml/texar, gaining over 2360 stars), an open-source machine learning and text generation toolkit based on TensorFlow.
- Designed and developed Forte (https://github.com/asyml/forte, gaining over 215 stars), a toolkit for building natural language processing pipelines, featuring cross-task interaction, adaptable data-model interfaces and many more (EMNLP 2020 System Demonstrations).

Research Intern December 2010 to May 2011

Internet Media Group,
Microsoft Research Asia, Beijing, China
Supervisor: Dr. Feng Wu and Dr. Chong Luo

- Analyzed the data collected from 54 sensors deployed in Intel Berkeley Research Lab to exploit the temporal correlations in sensor readings. Developed a joint source network coding scheme for approximate data gathering in wireless sensor network.

Skill Sets

- Proficiency with MATLAB, Python, Dynet, PyTorch, and TensorFlow
- Experienced in Java, R, C/C++, C#, AMPL

Honors and Awards

- North America Finalist of IBM Watson Build Challenge 2017
- Paper selected as the runner-up of the Best Paper in Electric Energy Systems Track of Hawaii International Conference on System Sciences 2015
- Founders Award of Excellence (top 1%) 2015
- Paper selected as one of the Best Conference Papers on Power System Analysis and Modeling of IEEE Power & Energy Society General Meeting 2014
- Excellent Graduate of Xidian University (top 1%) 2011
- National Scholarship (top 1%) 2010
- First prize of the College Academic and Technological Scholarship (top 2%) 2008-2010
- Excellent Student Awards (top 1%) 2008

Journal Publications

1. R. Wang, **P. Gao**, and M. Wang. "Robust Matrix Completion by Exploiting Dynamic Low-dimensional Structures." *submitted to EURASIP Journal on Advances in Signal Processing*, 2021. (The first two authors contributed equally.)
2. **P. Gao**, R. Wang, M. Wang, and J. H. Chow. "Low-rank Matrix Recovery from Noisy, Quantized and Erroneous Measurements." *IEEE Transactions on Signal Processing*, 2018, 66 (11): 2918-2932. (The first two authors contributed equally.)
3. **P. Gao**, M. Wang, J. H. Chow, M. Berger, and L. M. Seversky. "Missing Data Recovery for High-dimensional Signals with Nonlinear Low-dimensional Structures." *IEEE Transactions on Signal Processing*, 2017, 65 (20): 5421-5436.
4. **P. Gao**, M. Wang, J. H. Chow, S. G. Ghiocel, B. Fardanesh, G. Stefopoulos, and M. P. Razanousky. "Identification of Successive "Unobservable" Cyber Data Attacks in Power Systems Through Matrix Decomposition." *IEEE Transactions on Signal Processing*, 2016, 64 (21): 5557-5570.
5. **P. Gao**, M. Wang, S. G. Ghiocel, J. H. Chow, B. Fardanesh, and G. Stefopoulos. "Missing Data Recovery by Exploiting Low-dimensionality in Power System Synchrophasor Measurements." *IEEE Transactions on Power Systems*, 2016, 31 (2): 1006-1013.

Conference Publications

1. **P. Gao**, L. Zhang, Z. He, H. Wu, and H. Wang. "Improving Zero-shot Multilingual Neural Machine Translation by Leveraging Cross-lingual Consistency Regularization." *Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.
2. **P. Gao**, Z. He, H. Wu, and H. Wang. "Bi-SimCut: A Simple Strategy for Boosting Neural Machine Translation." *Proc. of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.
3. J. Li, **P. Gao**, X. Wu, Y. Feng, Z. He, H. Wu, and H. Wang. "Mixup Decoding for Diverse Machine Translation." *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.
4. R. Wang, T. Chen, Z. Xu, and **P. Gao**. "Robust Low-Rank Tensor Recovery From Quantized and Corrupted Measurements." *Proc. of Asilomar Conference on Signals, Systems, and Computers*, 2021.

5. Z. Liu, G. Ding, A. Bukkittu, M. Gupta, **P. Gao**, A. Ahmed, S. Zhang, X. Gao, S. Singhavi, L. Li, W. Wei, Z. Hu, H. Shi, X. Liang, T. Mitamura, E. Xing and Z. Hu. "A Data-Centric Framework for Composable NLP Workflows." *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

6. M. Wang, J. H. Chow, Y. Hao, S. Zhang, W. Li, R. Wang, **P. Gao**, C. Lackner, E. Farantatos, and M. Patel. "A Low-rank Framework of PMU Data Recovery and Event Identification." *Proc. of the First IEEE International Conference on Smart Grid Synchronized Measurements and Analytics (SGSMA)*, 2019.

7. G. Mijolla, S. Konstantinouplos, **P. Gao**, J. H. Chow, and M. Wang. "An Evaluation of Low-Rank Matrix Completion Algorithms for Synchrophasor Missing Data Recovery." *Proc. of the 20th Power Systems Computation Conference (PSCC)*, 2018.

8. **P. Gao**, and M. Wang. "Dynamic Matrix Recovery from Partially Observed and Erroneous Measurements." *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

9. M. Wang, J. H. Chow, **P. Gao**, Y. Hao, W. Li, and R. Wang. "Recent Results of PMU Data Analytics by Exploiting Low-dimensional Structures." *Proc. of the 10th Bulk Power Systems Dynamics and Control Symposium (IREP)*, 2017.

10. **P. Gao**, R. Wang, and M. Wang. "Quantized Low-rank Matrix Recovery with Erroneous Measurements: Application to Data Privacy in Power Grids." *Proc. of Asilomar Conference on Signals, Systems, and Computers*, 2016.

11. **P. Gao**, M. Wang, and J. H. Chow. "Matrix Completion with Columns in Union and Sums of Subspaces." *Proc. of IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2015.

12. M. Wang, J. H. Chow, **P. Gao**, X. T. Jiang, Y. Xia, S. G. Ghiocel, B. Fardanesh, G. Stefopoulos, Y. Kokai, N. Saito, and M. P. Razanousky. "A Low-Rank Matrix approach for the Analysis of Large Amounts of Synchrophasor Data." *Proc. of Hawaii International Conference on System Sciences (Runner-up of Best Paper in Electric Energy Systems Track)*, 2015.

13. M. Wang, **P. Gao**, S. G. Ghiocel, J. H. Chow, B. Fardanesh, G. Stefopoulos, and M. P. Razanousky. "Identification of "Unobservable" Cyber Data Attacks on Power Grids." *Proc. of IEEE SmartGridComm*, 2014.

14. **P. Gao**, M. Wang, S. G. Ghiocel, and J. H. Chow. "Modeless Reconstruction of Missing Synchrophasor Measurements." *Proc. of IEEE Power & Energy Society General Meeting (selected in Best Conference Paper sessions)*, 2014.

Technical Reports

1. Z. Hu, **P. Gao**, A. Bukkittu, and Z. Hu. "Introducing Texar-PyTorch: An ML Library integrating the best of TensorFlow into PyTorch." Octorber, 2019.

Patents

1. **P. Gao**, Z. He, Z. Li, and H. Wu. "A model training method, device, electronic equipment and storage medium." Application No.: CN202210186688.1, Filed February 28, 2022.

2. **P. Gao**, Z. He, Z. Li, and H. Wu. "Training method, text data processing method and device of deep learning model." Application No.: CN202210189268.9, Filed February 28, 2022.

3. **P. Gao**, Z. He, H. Wu, and H. Wang. "Methods, devices, equipment, media and program products for training models." Application No.: CN202111288550.4, Filed November 2, 2021.

4. X. Wan, J. Zhao, M. Wang, Z. He, H. Wu, Z. Li, Z. Xu, J. Liu, **P. Gao**, M. Sun, C. Li, and W. Yao. "Training method, device, electronic equipment and storage medium of translation model." Application No.: CN202111014476.7, Filed August 31, 2021.

5. J. Li, **P. Gao**, Z. He, and Z. Li. "Text translation device, electronic medium and method." Application No.: CN202110736794.8, Filed June 30, 2021.

6. J. Li, **P. Gao**, Z. He, and Z. Li. "Corpus generation method, device, electronic device and storage medium." Application No.: CN202110748376.0, Filed June 30, 2021.

7. **P. Gao**, Z. He, H. Wu, and H. Wang. "Model training method, device, electronic equipment and computer-readable storage medium." Application No.: CN202110320138.X, Filed March 25, 2021.

8. M. Wang, **P. Gao**, and J. H. Chow. "A low-rank-based missing PMU data recovery method." Application No.: 62/445305, Filed January 12, 2017.

Projects

Forte: A Data-Centric Framework for Composable NLP Workflows
Machine Learning Team,
Petuum, Inc., Pittsburgh, PA, USA

- Forte is a toolkit for building NLP pipelines, featuring composable components, convenient data interfaces, and cross-task interaction. Forte designs a universal data representation format for text, making it a one-stop platform to assemble state-of-the-art NLP/ML technologies, ranging from Information Retrieval, Natural Language Understanding to Natural Language Generation. Forte was originally developed in CMU and is actively contributed by Petuum in collaboration with other institutes. I am one of the main contributors of Forte repository.

Texar-PyTorch: A Modularized, Versatile, and Extensible Toolkit for Text Generation
Machine Learning Team,
Petuum, Inc., Pittsburgh, PA, USA

- Texar-PyTorch is a toolkit aiming to support a broad set of machine learning, especially natural language processing and text generation tasks. Texar-PyTorch provides a library of easy-to-use ML modules and functionalities for composing whatever models and algorithms. The tool is designed for both researchers and practitioners for fast prototyping and experimentation. Texar-PyTorch was originally developed and is actively contributed by Petuum and CMU in collaboration with other institutes. I am one of the main contributors of Texar-PyTorch repository.

Professional Activities & Service

- Student Member of IEEE, 2013 - 2017. Member of IEEE, 2018 - present.
- RPI Student Representative at the Center for Ultra-wide-area Resilient Electric Energy Transmission Networks (CURENT), 2015 - 2016.
- Program Committee Member:
  Conference on Uncertainty in Artificial Intelligence (UAI) 2018,
  The Annual Meeting of the Association for Computational Linguistics (ACL) 2023,
  The ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2023,
  The China National Conference on Computational Linguistics (CCL) 2023.
- Reviewer:
  IEEE Transactions on Signal Processing,
  IEEE Transactions on Smart Grid,
  IEEE Transactions on Automatic Control,
  IEEE Transactions on Power Delivery,
  IEEE/ACM Transactions on Networking,
  IEEE Signal Processing Letters,
  Annals of Mathematics and Artificial Intelligence,
  The International Conference on Computational Linguistics (COLING),
  The China National Conference on Computational Linguistics (CCL),
  American Control Conference (ACC),
  IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm),
  Intelligent System Applications to Power Systems (ISAP) Conference.