Pengzhi Gao

CONTACT INFORMATION

Xiaomi Campus
No.33 Xierqi Middle Road,
Haidian District, Beijing, China 100085

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

University of Pennsylvania, Philadelphia, PA, USA

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

Xidian University, Xi'an, China

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

WORK EXPERIENCE

Senior Algorithm Engineer & Tech Lead Manager June 2024 to Present

Large Language Model Team,
Xiaomi, Inc., Beijing, China
Supervisor: Wei Liu and Jian Luan

- Led the UI Agent team of ten talented researchers and engineers to develop UI Agent models that facilitate end-to-end language-to-mobile actions.
- Developed many-to-many multilingual machine translation systems based on large language models such as LLaMA-3, Qwen2.5, and Gemma2 models, which support real-time subtitle translation in XiaoAi Translate.

Staff Research & Development Engineer September 2020 to June 2024

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

- Developed a one-for-all multilingual sentence representation model, MuSR, that supports more than 220 languages (EMNLP 2023 Industry Track), which is used as the foundation model for the similarity-based bitext mining system in Baidu Translate.
- Proposed a simple but effective training strategy to boost the zero-shot performance of multilingual neural machine translation (ACL 2023 Findings). Developed data processing and model training pipelines 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.

Data Scientist / Machine Learning Engineer February 2018 to April 2020

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

- Designed and implemented the machine learning library for Petuum AI Builder Platform.
- Designed and developed Texar-PyTorch (gaining over 740 stars on GitHub), an open-source machine learning and text generation toolkit based on PyTorch. Maintained and contributed to Texar (gaining over 2380 stars on GitHub), an open-source machine learning and text generation toolkit based on TensorFlow.
- Designed and developed Forte (gaining over 240 stars on GitHub), 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 the wireless sensor network.

Skills

- **Research**: Machine Learning, Generative AI, Natural Language Processing, Signal Processing, High-dimensional Statistics
- **Programming**: Python > C/C++ = Matlab > Java = R
- **Software**: PyTorch, TensorFlow, Keras, Scikit-learn, Linux, MacOS, Git

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

Preprints & Technical Reports

1. **P. Gao**, Z. He, H. Wu, and H. Wang. "Towards Boosting Many-to-Many Multilingual Machine Translation with Large Language Models." *arXiv:2401.05861*, 2024.
2. R. Wang, **P. Gao**, and M. Wang. "Robust Matrix Completion by Exploiting Dynamic Low-dimensional Structures." *DOI: 10.21203/rs.3.rs-420556/v1*, 2021. (The first two authors contributed equally.)
3. Z. Hu, **P. Gao**, A. Bukkittu, and Z. Hu. "Introducing Texar-PyTorch: An ML Library integrating the best of TensorFlow into PyTorch." October, 2019.

Journal Publications

1. **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.)
2. **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.
3. **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.
4. **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. M. Cui, **P. Gao**, W. Liu, J. Luan, and B. Wang, "Multilingual Machine Translation with Open Large Language Models at Practical Scale: An Empirical Study." Proc. of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL), 2025. (The first two authors contributed equally.)

2. M. Sun, W. Liu, J. Luan, **P. Gao**, and B. Wang. "Mixture of Diverse Size Experts." *Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP): Industry Track*, 2024.

3. J. Du, Y. Wang, W. Zhao, Z. Deng, S. Liu, R. Lou, H. P. Zou, P. N. Venkit, N. Zhang, M. Srinath, H. R. Zhang, V. Gupta, Y. Li, T. Li, F. Wang, Q. Liu, T. Liu, **P. Gao**, C. Xia, C. Xing, J. Cheng, Z. Wang, Y. Su, R. S. Shah, R. Guo, J. Gu, H. Li, K. Wei, Z. Wang, L. Cheng, S. Ranathunga, M. Fang, J. Fu, F. Liu, R. Huang, E. Blanco, Y. Cao, R. Zhang, P. S. Yu, and W. Yin. "LLMs assist NLP Researchers: Critique Paper (Meta-)Reviewing." *Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

4. **P. Gao**, R. Zhang, Z. He, H. Wu, and H. Wang. "An Empirical Study of Consistency Regularization for End-to-End Speech-to-Text Translation." *Proc. of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

5. **P. Gao**, L. Zhang, Z. He, H. Wu, and H. Wang. "Learning Multilingual Sentence Representations with Cross-lingual Consistency Regularization." *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP): Industry Track*, 2023.

6. **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.

7. **P. Gao**, Z. He, H. Wu, and H. Wang. "Bi-SimCut: A Simple Strategy for Boosting Neural Machine Translation." *Proc. of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

8. 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.

9. 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.

10. 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): System Demonstrations*, 2020.

11. 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.

12. 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.

13. **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.

14. 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.

15. **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.

16. **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.

17. 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.

18. 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.

19. **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.

Patents

1. **P. Gao**, Z. He, and H. Wu. "用于多语言翻译的语言模型的训练方法及翻译方法." Application No.: CN202311759677.9, Filed December 20, 2023.

2. **P. Gao**, R. Zhang, Z. He, and H. Wu. "语音翻译模型的训练方法、语音翻译方法和装置." Application No.: CN202311622229.4, Filed November 30, 2023.

3. **P. Gao**, L. Zhang, Z. He, Z. Li, and H. Wu. "多语言的句向量的获取方法和多语言的编码器的训练方法." Application No.: CN202310395274.4, Filed April 13, 2023.

4. **P. Gao**, Z. He, Z. Li, and H. Wu. "深層学習モデルのトレーニング方法及び装置、テキストデータ処理方法及び装置、電子機器、記憶媒体、並びにコンピュータプログラム." Application No.: JP2022-190230, Filed November 29, 2022.

5. **P. Gao**, Z. He, Z. Li, and H. Wu. "Method of training deep learning model and method of processing text data." Application No.: US18/059,389, Filed November 28, 2022.

6. **P. Gao**, Z. He, Z. Li, and H. Wu. "深度学习模型的训练方法、文本数据处理方法和装置." Application No.: CN202210189268.9, Filed February 28, 2022.

7. **P. Gao**, Z. He, Z. Li, and H. Wu. "一种模型训练方法、装置、电子设备及存储介质." Application No.: CN202210186688.1, Filed February 28, 2022.

8. **P. Gao**, Z. He, H. Wu, and H. Wang. "用于训练模型的方法、装置、设备、介质和程序产品." Application No.: CN202111288550.4, Filed November 2, 2021.

9. 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. "翻译模型的训练方法、装置、电子设备及存储介质." Application No.: CN202111014476.7, Filed August 31, 2021.

10. J. Li, **P. Gao**, Z. He, and Z. Li. "文本翻译方法、装置、电子设备及存储介质." Application No.: CN202110736794.8, Filed June 30, 2021.

11. J. Li, **P. Gao**, Z. He, and Z. Li. "语料生成方法、装置、电子设备以及存储介质." Application No.: CN202110748376.0, Filed June 30, 2021.

12. **P. Gao**, Z. He, H. Wu, and H. Wang. "模型训练方法、装置、电子设备及计算机可读存储介质." Application No.: CN202110320138.X, Filed March 25, 2021.

13. 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 and Natural Language Understanding to Natural Language Generation. Forte was originally developed at CMU and is actively contributed by Petuum in collaboration with other institutes. I am one of the main contributors to Forte.

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 to Texar-PyTorch.

Professional Activities & Service

- Member of IEEE, 2018 - present. Member of ACL, 2022 - 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),
  The Annual AAAI Conference on Artificial Intelligence (AAAI),
  The Annual Meeting of the Association for Computational Linguistics (ACL),
  The Conference on Empirical Methods in Natural Language Processing (EMNLP) Industry Track,
  The Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING),
  The ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD),
  SIAM International Conference on Data Mining (SDM),
  The China National Conference on Computational Linguistics (CCL).
- Journal Reviewer:
  IEEE Transactions on Signal Processing,
  IEEE Transactions on Audio, Speech and Language 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,
  Journal of Data-centric Machine Learning Research (DMLR).
- Conference Reviewer:
  The Conference on Neural Information Processing Systems (NeurIPS),
  The International Conference on Learning Representations (ICLR),
  The Association for Computational Linguistics Rolling Review (ACL ARR),
  The Annual Meeting of the Association for Computational Linguistics (ACL),
  The Conference on Empirical Methods in Natural Language Processing (EMNLP),
  The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL),
  The Conference of the European Chapter of the Association for Computational Linguistics (EACL),
  The International Conference on Computational Linguistics (COLING),
  The China National Conference on Computational Linguistics (CCL),

The CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC),
The ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD),
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
American Control Conference (ACC),
IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm),
Intelligent System Applications to Power Systems Conference (ISAP),
International Conference on Intelligent Multilingual Information Processing (IMLIP).