高鹏至

手机：(+86) 15810512592 · **邮箱：**gpengzhi@gmail.com
个人主页：https://gpengzhi.github.io/
地址：北京市海淀区西二旗中路 33 号小米科技园

教育背景

伦斯勒理工学院，特洛伊，纽约州，美国 2013 年 8 月 - 2017 年 12 月
工学博士，电子工程

宾夕法尼亚大学，费城，宾夕法尼亚州，美国 2011 年 9 月 - 2013 年 5 月
工学硕士，电子工程

西安电子科技大学，西安，中国 2007 年 8 月 - 2011 年 7 月
工学学士 (优秀毕业生)，电子信息工程

工作经历

高级算法工程师 & 技术经理 2024 年 6 月 - 至今

大模型团队，**小米**，北京，中国

- 任职小米 UI Agent 团队负责人，领导二十余位人工智能研究人员、工程师和实习生开发基于多模态大模型的 GUI 智能体解决方案，以实现端到端的语言到手机操作执行的流畅体验。
- 训练多语言翻译大模型 GemmaX2 (NAACL 2025)，在 HuggingFace 机器翻译模型类别取得 Trending 第一并获得超过 10 万的下载量，提升小爱翻译效果并支持 28 个语种的互译。

资深研发工程师 2020 年 9 月 - 2024 年 6 月

自然语言处理部，**百度**，北京，中国

- 训练支持 220+ 语种的多语言句向量模型 (EMNLP 2023 Industry Track)，该模型被用于百度翻译双语数据挖掘系统中并持续产出高质量平行语料。
- 提出一种有效提升多语言机器翻译零资源方向效果的训练机制（ACL 2023 Findings），基于该策略开发多语言模型训练流程，提升百度多语言翻译系统的翻译效果，支持百度翻译 200+ 语种互译。
- 提出一种简单有效地提升机器翻译效果的训练机制（NAACL 2022），基于该策略开发模型训练流程，提升百度翻译模型的翻译效果。

数据科学家 / 机器学习工程师 2018 年 2 月 - 2020 年 4 月

机器学习团队，**Petuum**，匹兹堡，宾夕法尼亚州，美国

- 设计并开发用于 Petuum 人工智能开发平台的机器学习算法库 (基于 TensorFlow, Dynet 和 LightGBM)。
- 设计并开发多个著名开源项目：基于 TensorFlow 的机器学习与文本生成工具 Texar（在 GitHub 上获得超过 2380 个星标)；基于 PyTorch 的机器学习与文本生成工具 Texar-PyTorch（在 GitHub 上获得超过 740 个星标)；用于文本处理的自然语言处理流水线工具 Forte（在 GitHub 上获得超过 240 个星标)。

研究实习生 2010 年 12 月 - 2011 年 5 月

网络多媒体组，**微软亚洲研究院**，北京，中国

- 分析采集于 Intel-Berkeley 实验室的 54 个传感器的数据，探究其时域关联性。提出并开发了一种用于无线传感器网络数据采集的联合来源网络编码机制。

技术能力

- 研究领域：生成式 AI，智能体，机器学习，自然语言处理，信号处理，高维统计
- 编程语言：Python > C/C++ = Matlab > Java = R
- 软件：PyTorch, TensorFlow, Keras, Scikit-learn, Linux, MacOS, Git

获奖情况

- IBM Watson Build Challenge 北美总决赛入围者 2017 年
- 论文被选为 runner-up of the Best Paper in Electric Energy Systems Track of Hawaii International Conference on System Sciences 2015 年
- Founders Award of Excellence (前 1%) 2015 年
- 论文被选为 one of the Best Conference Papers on Power System Analysis and Modeling of IEEE Power & Energy Society General Meeting 2014 年
- 西安电子科技大学优秀毕业生 (前 1%) 2011 年
- 国家奖学金 (前 1%) 2010 年
- 西安电子科技大学一等奖学金 (前 2%) 2008 年 - 2010 年
- 西安电子科技大学学习标兵 (前 1%) 2008 年

预印论文 & 技术报告

- G. Liu, J. Ye, J. Liu, Y. Li, W. Liu, **P. Gao**, J. Luan, and Y. Liu. "Hijacking JARVIS: Benchmarking Mobile GUI Agents against Unprivileged Third Parties." *arXiv: 2507.04227*, 2025.
- Q. Wu, **P. Gao**, W. Liu, and J. Luan. "BacktrackAgent: Enhancing GUI Agent with Error Detection and Backtracking Mechanism." *arXiv: 2505.20660*, 2025.
- W. Xu, Z. Jiang, Y. Liu, **P. Gao**, W. Liu, J. Luan, Y. Li, Y. Liu, B. Wang, and B. An. "Mobile-Bench-v2: A More Realistic and Comprehensive Benchmark for VLM-based Mobile Agents." *arXiv: 2505.11891*, 2025.
- K. Huang, W. Xu, Y. Liu, Q. Wang, **P. Gao**, W. Liu, J. Luan, B. Wang, and B. An. "Enhance Mobile Agents Thinking Process via Iterative Preference Learning." *arXiv: 2505.12299*, 2025.
- **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.
- 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.)
- Z. Hu, **P. Gao**, A. Bukkittu, and Z. Hu. "Introducing Texar-PyTorch: An ML Library integrating the best of TensorFlow into PyTorch." Octorber, 2019.

期刊论文

- **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.)
- **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.
- **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.
- **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.

会议论文

- 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.)

- 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.
- 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.
- **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.
- **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.
- **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.
- **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.
- 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.
- 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.
- 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. P. Xing, and Z. Hu. "A Low-rank Framework of PMU Data Recovery and Event Identification." *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, 2020.
- 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.
- 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.
- **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.
- 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.
- **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.
- **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.
- 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.
- 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.
- **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.

专利

- **P. Gao**, R. Zhang, Z. He, and H. Wu. "Method and device for training speech translation model, and storage medium." Application No.: US18/930,081, Filed October 29, 2024.
- **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.
- **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.
- **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.
- **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.
- **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.
- **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.
- **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.
- **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.
- 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.
- 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.
- 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.
- **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.
- 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.

项目经历

Forte: A Data-Centric Framework for Composable NLP Workflows

机器学习团队，**Petuum**，匹兹堡，宾夕法尼亚州，美国

- Forte 是一个用于构建自然语言处理（NLP）流水线的工具包，具有可组合组件、便捷的数据接口和跨任务交互的特点。Forte 为文本设计了一种通用的数据表示格式，使其成为一个集最新 NLP/ML 技术于一体的一站式平台，涵盖信息检索、自然语言理解到自然语言生成等多个领域。Forte 最初由卡内基梅隆大学（CMU）开发，现由 Petuum 与其他机构合作积极维护。我是 Forte 的主要贡献者之一。

Texar-PyTorch: A Modularized, Versatile, and Extensible Toolkit for Text Generation

机器学习团队，**Petuum**，匹兹堡，宾夕法尼亚州，美国

- Texar-PyTorch 是一个旨在支持广泛机器学习任务，特别是自然语言处理和文本生成任务的工具包。Texar-PyTorch 提供了一个易于使用的机器学习模块和功能库，用于构建各种模型和算法。该工具旨在满足研究者和从业者快速原型开发和实验的需求。Texar-PyTorch 最初由 Petuum 和卡内基梅隆大学（CMU）联合其他机构开发。我是 Texar-PyTorch 的主要贡献者之一。

专业活动

- IEEE 会员，2018 年 - 至今。ACL 会员，2022 年 - 至今。
- 领域主席：
  The Association for Computational Linguistics Rolling Review (ACL ARR).
- 期刊审稿人：
  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).

- 程序委员会成员 / 会议审稿人:
  The Conference on Uncertainty in Artificial Intelligence (UAI),
  The Annual AAAI Conference on Artificial Intelligence (AAAI),
  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 Conference on Language Modeling (COLM),
  The China National Conference on Computational Linguistics (CCL),
  The CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC),
  SIAM International Conference on Data Mining (SDM),
  The ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD),
  IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
  IEEE/CVF International Conference on Computer Vision (ICCV),
  IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm),
  American Control Conference (ACC),
  Intelligent System Applications to Power Systems Conference (ISAP),
  International Conference on Intelligent Multilingual Information Processing (IMLIP).