# 高鹏至

**手机：**(+86) 15810512592 · **邮箱：**gpengzhi@gmail.com

**个人主页：**https://gpengzhi.github.io/

**地址：北京市海淀区西北旺东路 10 号院百度科技园 1 号楼**

## 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

- 导师：汪孟教授
- 博士论文：High-dimensional Data Analysis by Exploiting Low-dimensional Models with Applications in Synchrophasor Data Analysis in Power Systems

**宾夕法尼亚大学**，费城，宾夕法尼亚州，美国 2011 年 8 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 5 月

*工学学士* (优秀毕业生)，电子信息工程

## 研究方向

我的研究方向包括机器学习，自然语言处理，信号处理和高维数据分析。我特别关注低维模型与优化算法在电网分析、图像处理与机器翻译中的应用。

## 工作经历

**百度自然语言处理部**，北京，中国，**资深研发工程师** 2020 年 9 月 - 至今

- 维护并优化多语言机器翻译系统，支持百度翻译 200+ 语种互译。
- 提出一种用于机器翻译多样性的解码机制（EMNLP 2021 Findings），并基于该机制开发数据增强方法，以提升百度翻译模型的翻译效果。
- 提出一种有效提升机器翻译效果的训练机制（NAACL 2022），并基于该策略开发模型训练流程，以提升百度翻译模型的翻译效果。

**Petuum, Inc.**，匹兹堡，宾夕法尼亚州，美国，**数据科学家** 2018 年 2 月 - 2020 年 4 月

- 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)。
- 设计并开发基于 PyTorch 的机器学习与文本生成工具箱 Texar-PyTorch（https://github.com/asyml/texar-pytorch），在 GitHub 上获得超过 705 stars。
- 开发并维护基于 TensorFlow 的机器学习与文本生成工具箱 Texar（https://github.com/asyml/texar），在 GitHub 上获得超过 2250 stars。
- 设计并开发用于文本处理的自然语言处理流水线工具 Forte（https://github.com/asyml/forte），在 GitHub 上获得超过 150 stars（EMNLP 2020 System Demonstrations）。

**微软亚洲研究院**，北京，中国，**研究实习生** 2010 年 12 月 - 2011 年 5 月

- 分析采集于 Intel-Berkeley 实验室的 54 个传感器的数据，并探究其时域关联性。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

- 熟练：MATLAB，Python，Dynet，PyTorch，TensorFlow
- 有经验：Java，R，C/C++，C#，AMPL

## 获奖情况

- IBM Watson Build Challenge 北美决赛入围者 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

## 期刊论文

- 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.)
- **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

- **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.

## 技术报告

- Zecong Hu, **Pengzhi Gao**, Avinash Bukkittu, and Zhiting Hu. "Introducing Texar-PyTorch: An ML Library integrating the best of TensorFlow into PyTorch." Octorber, 2019.

## 专利

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

- Forte 是一个用于构建自然语言处理任务流程的工具包，其包含了可组合组件、方便的数据接口和跨任务交互。Forte 为文本设计了一种通用数据表示格式，使其成为汇集从信息检索、自然语言理解到自然语言生成等最新 NLP/ML 技术的一站式平台。Forte 最初由卡内基梅隆大学开发，并由 Petuum 与其他机构合作积极贡献开发。我是 Forte 代码库的主要贡献者之一。

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

- Texar-PyTorch 是一个工具包，旨在支持广泛的机器学习任务，尤其是自然语言处理和文本生成任务。Texar-PyTorch 提供了一个易于使用的机器学习模块和功能库，用于组合任何模型和算法。该工具是

为研究人员和从业者设计的，用于快速迭代原型和实验。Texar-PyTorch 最初由 Petuum 和卡内基梅隆大学与其他研究所合作开发。我是 Texar-PyTorch 代码库的主要贡献者之一。

**DyNet: The Dynamic Neural Network Toolkit**

- DyNet 是由卡内基梅隆大学、Petuum 和许多其他公司开发的神经网络库。它是用 C++ 编写的（用 Python 绑定），它被设计成在 CPU 或 GPU 上运行时是高效的，并且能很好地处理具有动态结构的网络，每一个训练实例动态结构都会发生变化。我贡献了 DyNet 代码库的示例和教程部分。

## 专业活动

- IEEE 学生会员，2013 年 - 2017 年。IEEE 会员，2018 年 - 至今。
- Center for Ultra-wide-area Resilient Electric Energy Transmission Networks (CURENT) 伦斯勒理工学院学生代表。
- 教学助理（伦斯勒理工学院）:
  Modeling and Analysis of Uncertainty，2017 年秋季学期,
  Distributed Systems and Sensor Networks，2017 年秋季学期。
- 程序委员会成员:
  Conference on Uncertainty in Artificial Intelligence (UAI) 2018。
- 审稿人:
  IEEE Transactions on Signal Processing,
  IEEE Transactions on Smart Grid，
  IEEE Transactions on Automatic Control，
  IEEE/ACM Transactions on Networking，
  IEEE Signal Processing Letters，
  Annals of Mathematics and Artificial Intelligence，
  American Control Conference，
  IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)，
  International Symposium on Antennas and Propagation。