Guoxing Lan 兰 国兴

Cell phone: +86-18810456462 ☎
Email: langx15@tsinghua.org.cn ✉
Github: https://github.com/lankuohsing 🔗
Tech Blog: https://blog.csdn.net/thuchina 🔗
Homepage: https://lankuohsing.github.io/ 🔗

Research Interests:

Natural Language Processing, Representation Learning, Acceleration of Machine Learning Models, and Continual Learning.

Education Experiences:

- **Department of Automation, Tsinghua University** *Sep. 2015 – July 2018*
  M. Eng. in Control Science and Engineering. Supervisor: Prof. Nong Cheng and Prof. Qing Li
- **School of Economics and Management, Tsinghua University** *Aug. 2013 – July 2015*
  B. Ec. in Economics (For Second Bachelor Degree). Excellent graduation thesis
- **Department of Automation, Tsinghua University** *Aug. 2011 – July 2015*
  B. Eng. in Automation

Industry Experiences:

- **Huawei Consumer Business Group** *Algorithm Engineer, Senior Algorithm Engineer* *Aug. 2018 – Present*
  I'm currently doing pre-research on the miniaturization, acceleration and personalization of AI applications;
  I implemented machine learning algorithms for the NLU system of Huawei's voice assistant (Xiaoyi);
  I have applied for 6 patents: one for task decision in dialogue system (published), two for fast algorithms for NLP applications, two for model compression algorithms for AI applications, and one for model fine-tuning algorithms for personalized AI applications.
- **JD Group** *Algorithm Engineer Intern* *Aug. 2017 – Nov. 2017*
  I focused on optical character recognition algorithms and applications for train tickets.

Research Experiences:

- **Department of Automation , Tsinghua University** *Sep. 2016 – Feb. 2018*
  Research Assistant; Supervisor: Prof. Nong Cheng and Prof. Qing Li
  I was mainly engaged in the research on modeling and simulation of turbofan engine, data-driven fault diagnosis and remaining useful life estimation.
- **Department of Electrical and Computer Engineering, University of Alberta** *July. 2014 – Sep. 2014*
  Research Intern; supervisor: Prof. Venkata Dinavahi
  I designed and implemented PSO algorithms for nonlinear constrained optimization problems.

Selected Projects:

- **Personalized ASR Algorithms and Application** *Aug. 2021 –present*
  I lead a team of 4 members conducting pre-research on personalized ASR algorithms and application. We design and implement an architecture to fine-tune a pre-trained ASR model for each accented user. Our goal is to improve the performance of target users without reducing the performance of ordinary users. My main responsibility is to design the entire architecture and experimental plans, and implement some key modules such as the main program and model evaluation module. During this project, I proposed **a novel method to alleviate the problem of knowledge forgetting** when fine-tuning classification models (with a softmax layer) with new personalized samples. The core idea is to select "important samples" in the original training set, mix them with the new personalized samples, and compose the training set for fine-tuning. The "important samples" are those original training samples located near the linear decision boundary of the softmax layer. A relevant patent has been applied for.

- **Miniaturization and Acceleration of AI models** *July. 2020 –July. 2021*

  I focused on the miniaturization and acceleration of AI models to reduce the consumption of ROM, RAM and power, and increase the inference speed, with little or no accuracy loss. My main contributions: I **optimized the code of GPU operators in Tensorflow-Lite** to reduce their inference latency; I proposed **a new encoding method for quantized weights of AI models**, reducing the consumption of storage and network transmission; I proposed a **hierarchical softmax layer with a novel method for the arrangement of leaf nodes** for text generation models, increasing the inference speed with little accuracy loss.

- **NLU of Huawei's Voice Assistant** *Sep. 2018 –July. 2019*

  The NLU (Natural Language Understanding) system of Huawei's voice assistant (Xiaoyi) follows a hierarchical structure: domain classification, intent detection, slot filling and task decision-making. I was responsible for the domain classification module and task decision-making module. To support the rapid expansion of Xiaoyi's business and the quick repairing of Xiaoyi's online bugs, the input space of domain classification is divided into 20 domains (classes), and each domain is trained with an SVM classifier based on LIBSVM. My main contributions include: I used unigram and bigram to vectorize the input text and extract its TF-IDF features; I used the information-gain method to select important features to avoid the dimensional disaster problem; **I optimized the inference code of LIBSVM and got an inference latency reduction of 10~20ms; I added regex and custom dictionaries to the feature vector to help fix bugs and quickly support new hot words.** Besides, to improve the end-to-end accuracy, the NLU system triggers multiples intents (each intent corresponds to a task), and I designed a re-ranking module to score the candidates and choose the top-1 intent as the final output. Xiaoyi has served hundreds of millions of smart phone users.

- **Remaining Useful Life Estimation Based on LSTM** *Nov. 2017 –Feb. 2018*

  The aim of this project is to estimate the remaining useful life (RUL) of turbofan engine with degradation. I proposed **a novel data-driven method based on LSTM neural network** to estimate the RUL with multi-variable outputs of sensors and operational settings. Since the degradation of the engine is usually negligible, the RUL of the engine is a constant value in the early stage until the failure starts after the engine has been running for a period of time. I proposed a **novel method to identify the initial RUL**: split the total RUL into m "windows", calculate the geometric center of the sensors' outputs for each time stamp, calculate the geometric center of the formal geometric centers of all the time stamps in every window, calculate the Euclidean distance between each window's geometric center and that of the first window, plot all the Euclidean distances and finally treat the inflection point as the start of the failure. Experiments show that my method outperforms other methods on NASA's C-MAPSS dataset.
  Project link: https://github.com/lankuohsing/Remaining-Useful-Life-Prediction-RNN

Publications:

- **G. Lan**, Q. Li and N. Cheng, "Remaining Useful Life Estimation of Turbofan Engine Using LSTM Neural Networks," 2018 IEEE CSAA Guidance, Navigation and Control Conference (CGNCC), Xiamen, China, 2018, pp. 1-5. (Oral) [PDF]
- **G. Lan**, N. Cheng and Q. Li, "Comparison and fusion of various classification methods applied to aero-engine fault diagnosis," 2017 29th Chinese Control And Decision Conference (CCDC), Chongqing, 2017, pp. 4754-4759. (Oral, Session chair)[PDF]

Others:

Programming skills: Python, Java, C/C++, TensorFlow, Pytorch, Matlab, etc.
Knowledge: linear algebra, calculus, statistics, convex optimization, data structures and algorithms, foundations of NLP, foundations of machine learning/deep learning, etc.
Leisure interests: writing technology blogs, pushing codes to GitHub, and sharing knowledge; Physical exercise including strength and aerobic training.