Lin Wei

AI Model Optimization Engineer | Efficient LLM Specialist
San Diego, CA | (619) 555-3214 | lin.wei@example.com
linkedin.com/in/linwei-ai | github.com/linwei-efficient-llm

Professional Summary

Detail-oriented AI Model Optimization Engineer with 5+ years of experience specializing in efficiency techniques for large language models. Expert in quantization, pruning, and distillation methodologies. Proven track record of reducing model size and computation requirements while maintaining performance quality.

Professional Experience

ML Optimization Engineer

EfficientAI Labs, San Diego, CA (Nov 2020 - Present)
- Developed 4-bit quantization techniques reducing model size by 75% with less than 2% performance degradation. - Implemented structured pruning methods achieving 40% parameter reduction while maintaining 95% of baseline performance. - Created knowledge distillation pipeline for transferring capabilities from 70B to 7B parameter models. - Designed specialized architectures for mobile and edge deployment of language models.

Machine Learning Engineer

MobileCompute Inc., San Jose, CA (Mar 2018 - Oct 2020)
- Optimized neural networks for mobile and edge devices using TensorFlow Lite and ONNX. - Implemented quantization-aware training for computer vision and NLP models. - Developed benchmark suites for evaluating model efficiency across devices. - Created automated pipelines for model optimization experiments.

Software Engineer

TechOptimize, San Francisco, CA (Jul 2016 - Feb 2018)
- Developed performance optimization tools for mobile applications. - Implemented memory profiling and optimization techniques. - Created benchmarking frameworks for measuring application performance.

Technical Skills

- **Model Compression:** Quantization (INT8/4/2), Knowledge Distillation, Model Pruning
- **Programming Languages:** Python, C++, CUDA
- **ML Frameworks:** PyTorch, TensorFlow, ONNX Runtime
- **Optimization Tools:** bitsandbytes, GPTQ, AWQ, Intel NNCL
- **Mobile & Edge ML:** TensorFlow Lite, CoreML, PyTorch Mobile
- **Performance Analysis:** NSight Compute, PyTorch Profiler, Intel VTune
- **Deployment:** ONNX Runtime, TensorRT, OpenVINO

Education

Master of Science, Computer Engineering

University of California, San Diego, CA (Graduated: May 2016)
- Thesis: “Efficient Neural Network Architectures for Resource-Constrained Environments” - GPA: 3.95/4.0

Bachelor of Engineering, Computer Science
Tsinghua University, Beijing, China (Graduated: Jun 2014)
- Focus on algorithms and systems optimization - Graduated with honors

---

Technical Publications

- Wei, L., et al. (2022). "Mixed-Precision Quantization for Language Models." *ICLR.*
- Wei, L., et al. (2021). "Knowledge Distillation Strategies for Efficient NLP Models." *EMNLP.*
- Wei, L., et al. (2020). "Structured Pruning for Transformer Models." *NeurIPS Workshop on Efficient ML.*

Patents

- US Patent Application 17/852,391: "Method for Efficient Quantization of Neural Networks"
- US Patent 11,475,291: "System for Low-Bit Representation of Neural Networks"

Open Source Contributions

- Creator of "LLM-Optimizer" - open-source toolkit for LLM compression
- Contributor to bitsandbytes library, focusing on quantization algorithms
- Maintainer of benchmarking suite for evaluating LLM efficiency

---

Certifications

- NVIDIA Deep Learning Institute - Accelerated Computing with CUDA (2022)
- TensorFlow Developer Certificate (2020)

Languages: English (fluent), Mandarin Chinese (native)