Work History

Basic Information

Name: Sou Yoshihara (   )

Service	Account
GitHub	@sousquared
Instagram	@sousquared
Twitter	@sou_squared
Zenn	@sousquared

Personal Career Philosophy

Since childhood, I have loved drawing manga and illustrations so much that I once aspired to become a manga artist. With this background, I work with the desire to contribute to the creative and entertainment fields. My goal is to create products and works that bring joy to users.

Skills

Languages

- Japanese
  - Native
- English
  - Conversational
  - TOEIC 875 (Apr 2019)
  - TOEFL 85 (Jun 2017)

Programming Languages & Frameworks

- Python
- FastAPI
- Triton Inference Server
- JavaScript/TypeScript
- React
- Terraform

Cloud/Infrastructure

- GCP (Google Cloud Platform)
  - Cloud Run
  - Cloud Build
  - GKE (Google Kubernetes Engine)
  - Cloud Composer
  - etc.

Development Experience

- PoC and production implementation using machine learning models
- API development (FastAPI)
- Clean Architecture and DDD (Domain-Driven Design) oriented coding

Others

- KaiRA (Kyoto University AI Research Association) Operating Member
  - Organizing reading sessions on artificial intelligence
- JEES SoftBank AI Talent Development Scholarship (2020-2021)

Strengths

- Can handle everything from PoC to production development
- Proactively taking action based on my interests
- Strong presentation and explanation skills
- International network of friends around the world

Interests

- Social applications of AI technology
- Mechanisms of human cognitive abilities (neuroscience, psychology, behavioral economics)
- AI-powered art and entertainment

Work Experience

Career Summary

As an engineer, I have not only developed features but also considered which PoCs and new features would become competitive advantages for the service and contribute to KPI improvements such as revenue growth, working backwards from business strategies. In the New Business Development team, I worked on improving product logic to reduce designers' production work, API development, and model serving. Also, as a leader of an ML team of 4-5 members, I led strategic planning and research roadmaps for updating internal product search functionality and new features, achieving quantitative results such as doubling search feature usage rates. In the Auto Generation team, I participated in 4 auto-generation logic projects over the course of one year, achieving service-level success rates in 3 of them. For example, I achieved 90% + quality in banner ad image auto-generation for a certain pattern within approximately 2 months. I built high-speed experiment pipelines and annotation-based improvement workflows, balancing development speed with quality improvement. Also, as operating leader and member of DSOps training, I built a system where participating in training leads to actual projects in the field, contributing to organizational growth.

Aug 2018 - Aug 2020

Organization: Hitachi, Ltd. / Kyoto University Lab
Position: Student Researcher (Part-time)
Responsibilities

- Research and implementation of the latest machine learning papers
- Research and development of algorithms adaptable to modeling geospatial information
- Literature surveys on Graph Convolutional Networks and Relational Graph Convolutional Networks
- Implementation of parts of models devised by researchers
- Algorithm evaluation

Jan 2021

Organization: CyberAgent / Kiwami Prediction AI Division / Prediction Team
Position: ML/DS Intern
Responsibilities: Responsible for updating and accuracy evaluation of video ad score prediction models.

Apr 2022 ~ Mar 2025

Organization: CyberAgent / Kiwami AI Division / New Business Development Team
Position: ML Engineer
Responsibilities

I worked on improving product logic to reduce designers' production work, API development, and model serving.

Strategic Planning

- As a leader of the ML team, led strategic planning and roadmap creation for updating internal product search functionality and new features

Search Feature Improvement

- Formulated hypotheses from user usage history and improved functionality. Specifically, by adding multimodal search using images and text (e.g., searching with a material image + “premium feel”), improved the usability of image search and **successfully doubled search feature usage rates**. Also experimented and implemented logic to filter out poor-quality search results, **filtering out 40% of poor-quality search results** and improving the quality of search results displayed to users

Backend Development

- Also handled API development, mainly gaining experience in microservice development using Cloud Run

MLOps

- Under senior MLOps members, worked on building and operating basic inference servers and inference pipelines (using GKE, Triton, Cloud Batch, Apache Airflow on Cloud Composer, etc.). As a personal achievement, migrated the inference infrastructure from CPU to GPU. As a result, **achieved 6.28x faster inference speed compared to CPU**. Also reduced processing time by **62%** through preprocessing acceleration (including cache utilization)

Cost Optimization

- Cleaned up unused GKE and VMs, **contributing to approximately $2,500 monthly cost reduction**

Organizational Contribution

- As a leader of an ML team of 4-5 members, promoted PoC and production development of new features utilizing AI
- Also, as operating leader of DSOps training, planned new initiatives and built a system where participating in training leads to actual projects in the field (see article for details. It is written in Japanese.)

Apr 2025 ~ Present

Organization: CyberAgent / Kiwami AI Division / Auto Generation Team
Position: ML Engineer
Responsibilities

Ad Auto-Generation Logic Development

- Participated in 4 auto-generation logic projects over the course of one year, achieving service-level success rates in 3 of them. Specifically, these projects involved consolidating ad patterns proposed by creators into requirements, translating them to implementation level, and confirming success rates through human annotation. Collaborated with 1-2 engineers and 1-2 PMs, promoting speed-focused development

Project Results

- (Since specific logic details are confidential, I will describe based on success rates) Among the successful projects, in the 1st project, collaborated with 5 engineers and **achieved 74% success rate quality in auto-generation within 2 months** (this was quite high accuracy before NanoBananaPro was released)
- In the 2nd project, worked with 2 engineers on banner ad image auto-generation for a certain pattern and **successfully developed auto-generation logic with 90% success rate quality within 3 months**. In this project, I mainly handled prompt tuning for image generation and image filtering, and implementation of generation pipelines

- In the 3rd project, worked with 2 engineers on banner ad image auto-generation for a certain pattern and **successfully developed auto-generation logic with 90%+ quality within 2 months**. In this project, I handled prompt tuning for image filtering and clear verbalization of annotation criteria. In all projects, collaborated with ad creators on requirements definition, rapidly iterated through experiment implementation, annotation requests, and feature improvement PDCA cycles, conducting speed-focused development. Also handled API implementation when necessary

Organizational Contribution

- As an operating member of DSOps training, realized a special lecture by Yuta Saito (Ph.D., Cornell University, Hanjuku Virtual Co., Ltd.). (I was the moderator for this article.)

Research

- Sou Yoshihara, Taiki Fukiage, Shin'ya Nishida, "Towards acquisition of shape bias: Training convolutional neural networks with blurred images.", VSS, Poster session, 2021.
- Sou Yoshihara, Taiki Fukiage, Shin'ya Nishida, "Does training with blurred images bring convolutional neural networks closer to humans with respect to robust object recognition and internal representations?", Front. Psychol., Vol. 14,2023
- (Japanese) Sou Yoshihara, Taiki Fukiage, Shin'ya Nishida, "Shape Bias ", VISION, Vol. 33, No.1, 1-5, 2021. Best Presentation Award, Vision Society of Japan 2020 Summer Conference code

Articles

- ” ” —— 7 DSOps (Japanese)
- ML/DS | CyberAgent Way (Japanese) (I was the moderator for this article.)
- Codex MCP AI Coding : Codex 3 (Japanese)
- I also write mainly technical articles on Zenn: Zenn:@sousquared (Japanese)

Personal Projects

Vocavisual: Linking words with visuals, beyond your native language.

This project aims to directly connect images with words being learned, without relying on one's native language. For example, when learning the word "cat," instead of translating it to " " (cat in Japanese) and memorizing it, the goal is to associate it with an image of a cat. By directly connecting images with words being learned without using one's native language as an intermediary, I believe this can promote deeper understanding. The project is being conducted on the following Instagram account:

- Vocavisual Korean: https://www.instagram.com/vocavisual_korean/