Ian Pointer

ML/Deep Learning Researcher (Search/LLMs)

AI Research Lead @ Bookend.AI

ian@snappishproductions.com • @carsondial • 919-450-5242 • Cincinnati, OH, USA • github •

Skills

Python, PyTorch, Scala, Java, Ruby, SQL, Go, JavaScript, Apache Spark, Kubernetes, Docker, Apache Solr, Apache Hadoop, Apache Kafka, Apache Storm, Apache Beam, TensorFlow, Keras, Seldon Core, Milvus, Vespa, Apache Flink, Argo, Redis, AWS, Google Compute Platform

Experience

2023-Present

AI Research Lead for Bookend.AI

Developed anti-jailbreaking techniques capable of stopping 99.8% of suffix-based attacks on llama2-based models and an extensible red-teaming LLM-based agent system to probe LLMs for a variety of vulnerabilities including persuasion and inception techniques.

Created a RLAIF architecture to steer LLM responses towards customer preferences using DPO/KTO/OPRO/etc. techniques and synthetic data generation.

Trained a suite of finance models, beating BloombergGPT with models over 20x smaller in parameter count.

Added an evaluation suite for generating evaluation prompt metrics through customer engagment, as well as point- & list-wise comparisons of model output with both human and LLM-based juries and using a non-ELO based scoring system to allow for faster evaluation of new models against existing deployed models.

Implemented evolutional model merging for LoRA adapters using Simultaneous Perturbation Stochastic Approximation to maximise the performance of customer fine-tuning.

2018-2023

ML Architect / Senior Data Engineer for Lucidworks

Created and deployed a retrieval-augmented generation (RAG) search service which allowed customers to choose different LLM providers (e.g. OpenAI, open source models, etc) with reflection and other techniques to reduce hallucinations in answers.

Developed new approaches for text and image-based semantic searching using a variety of different model architectures (e.g. RNNs, CNNs, BERT, CLIP, etc) and embeddings, training models on CPU, GPUs, and TPUs.

Created a flexible open-domain metadata pipeline using CLIP and T5-based models for zero-shot and fine-tuned annotations of customer's image and text data.

Engineering lead for deep learning-based approach to handling zero search results, implementation and deployment of solution resulting in multi-million dollars of extra revenue from various different e-tailers.

Lead of data analytics team, optimizing Apache Spark jobs at scale, migrating infrastructure to Kubernetes, adding new workflow orchestration with Argo, MLOps work for deployment, instrumentation and faster inference speeds.

Committer for spark-solr open source project.

Wrote *Programming PyTorch for Deep Learning: Creating and Deploying Deep Learning Applications* for O'Reilly for developing Deep Learning applications on images, text, and audio, using cutting-edge techniques and architectures.

2018-2018 **Director of Partner Engineering for Kogentix, Inc**

Responsible for the team that integrates the machine learning application AMP with partner platforms (e.g. Microsoft Azure (Databricks & HDInsight), Google Cloud Platform (Dataproc), Amazon Web Services (EMR), Cloudera).

2016-2018 **Senior Solutions Architect / Engineering Manager for Kogentix, Inc** (Remote)

Worked on cloud-based data warehousing and analytical projects for multiple global financial companies, including migrating data and infrastructure off-premises and re-architecting legacy applications to handle dynamic workloads in Spark.

Deep Learning projects using PyTorch and Keras for image segmentation and recognition / classification, plus DL NLP for sentiment analysis and classification.

2013-2016 **Lead Consultant for Mammoth Data** (Durham, NC)

Worked on a wide variety of projects for clients large and small, including designing, implementing, and maintaining massive-scale AWS and Google Cloud Platform deployments, data pipelines involving Kafka, Storm, and Spark, along with analysis such as recommendation engines, natural language parsing, and anomaly detection.

Architected and developed an ETL application for a real-time advertisement bidding company using Apache Spark and Redshift for processing incoming data for further analysis. (5bn records / day)

Designed and implemented an orchestration engine in Ruby revolving around OpenStack/Git/Jenkins for a large email marketing company that sends over 250m emails per day. The engine was responsible for deploying the organization's applications across dev, stage, and production, complete with pluggable APIs for Docker support and alternate cloud providers, along with capabilities for online scaling and self-healing of application groups.

Designed and implemented a multi-region AWS cloud setup using CoreOS, Docker, and Ansible for a company allowing them to increase availability and scalability for their Rails/Redis/MongoDB application stack, serving millions of API requests per day. Profiling and improving Rails applications using Ruby and OS-level profiling tools.

2013 **DevOps Engineer for Reverbnation** (Durham, NC)

DevOps in a high-speed, web-scale Ruby-on-Rails environment (> 3.5m users with 1m unique logins per day).Worked on the migration from Rails 2.3 to Rails 3 and helped improve the transactional mail system to increase throughput and reduce spam rates as well as moving to a new cloud provider and diagnosing and improving Redis/MySQL performance issues.

2012-2013 **Systems Administrator for Rho, Inc** (Durham, NC)

Maintaining the company's fleet of CentOS Tomcat application servers running on VMware vSphere. Developed a prototype iOS questionnaire app for internal testing, visualizing data using d3.js. Rolling out automated deployment of servers with Puppet & vSphere.

2011-2012 **Developer for Open Software Integrators** (Durham, NC)

Part of a team contracting for various Fortune 500 companies all over the USA, specializing in web front-end development (including a Single-Sign-On project to support one of the largest video game releases of 2011) and DevOps, carrying out troubleshooting and development roles in technologies such as Spring, tcServer/Tomcat, RHEL/CentOS, and HTML5+JavaScript+jQuery.

2005-2011 **Administrator and Developer for d'Overbroeck's College Ltd** (Oxford, UK)

Managed a team of four people, responsible for over 500 users across six sites in Oxford. Developed various bespoke applications and databases, including a marketing reports system, printer usage tracking software, a ticket-request system, and a web-based academic grading and reporting application complete with PDF output and email notification

to parents, utilizing technologies such as HTML5/CSS3, JavaScript, Ruby on Rails, Sinatra, CouchDB, Redis, MySQL, and SQL Server.

2003-2005 **Technical Journalist** (Oxford, UK)

Writing articles for Linux Journal and Linux Magazine on the subject of coding with Mono and Emulation in Linux. Contributor to Open Source Press Books on the subject of DVD Authoring.

2002-2003 **TA for University of North Carolina at Chapel Hill** (Chapel Hill, NC)

TA in Computer Science department.

2000-2002 **System Administrator for Oxford Brookes University** (Oxford, UK)

Part of a four-person team responsible for over 15,000 users. Maintained Solaris/ Windows NT/Windows 2000/OpenVMS systems. Extended and maintained university's Perl-based webmail.

Education

1997-2000 *BSc(Hons) in Computer Science* (1st Class Degree with Honours, Manchester, United Kingdom).

Publications

2019 *Programming PyTorch for Deep Learning: Creating and Deploying Deep Learning Applications (O'Reilly)*
2016-Present *Infoworld*
2013 *Instant Zepto.js (Packt Press)*
2005 *Open Source Video (Open Source Press)*

---

ian@snappishproductions.com • @carsondial • 919-450-5242 • Cincinnati, OH, USA • github •