No GPUs? No Problem

Training ML Models Without Expensive Hardware

ALL THINGS AI 2026

SUHA MOKALLA
March 24, 2026

attention arc

AGENDA

→ Who am I?

→ Problem at hand!

→ Solution

→ Applications

→ Key Takeaways

attention arc

Who am I?...

Background

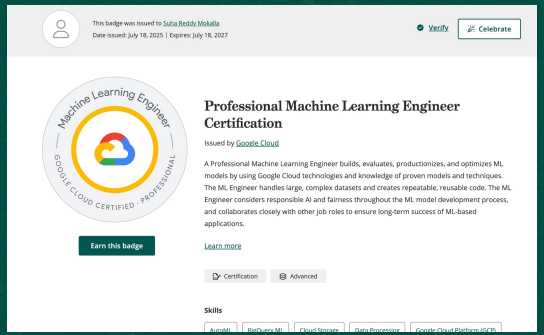
- Ph.D in Engineering, 2023
  The University of Georgia, Athens, GA
- M.S in Electrical Engineering, 2020
  West Virginia University, Morgantown, WV

Dissertation

- Face Detection and Recognition Technologies in visible and thermal spectra
- GAN (Generative Adversarial Network) Image Synthesis

Data Scientist

- Attention Arc
- Since Sep 2025
- Google Cloud Professional ML Engineer Certification



Key Projects

- Attribution Estimation for linear TV
- Propensity Score Modeling

attention arc

Model Training

→ Models need to be trained

→ Larger the model, bigger the GPU requirements

→ Stable Diffusion - requires NVIDIA RTX series GPU, with over 8GB of VRAM

→ LLMs - different specification require different GPUs

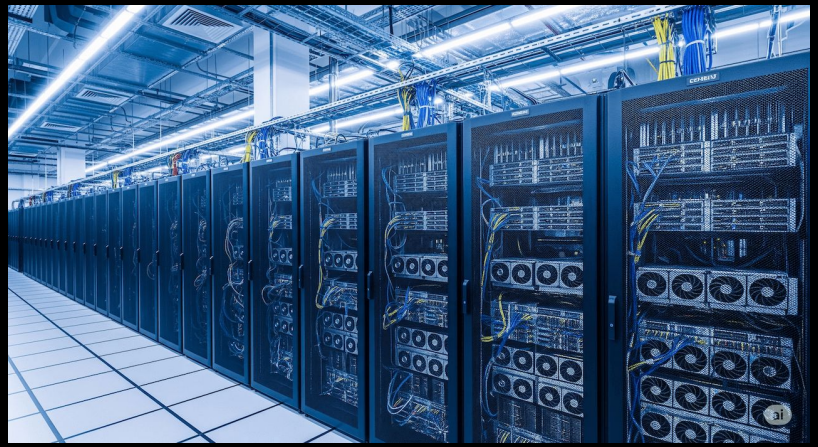
→ 7B-13B models need 12-24GB VRAM

→ 70B+ models might need multi-gpu clusters (4x or 8x A100/H100 80GB)

→ RAM is expensive

Problem

	Purchase Cost
RTX 4090	$1599
A100	$4699 (basic)
H100	~$30000



Expensive to purchase

A100 launched 2022 and discontinued 2024

Additional Cost - Maintenance, replacement, cooling, electric, setting up machines, physical space etc.

attention arc

Solution

→ Cloud resources - GCP Vertex AI, AWS Sagemaker, SnowflakeML

→ “Renting” a GPU when needed

→ No physical equipment maintenance

→ No space needed

→ Easy reproducibility

→ Accessible to multiple developers

→ Scale to any number of GPUs in an instant

attention arc

GCP - Vertex AI

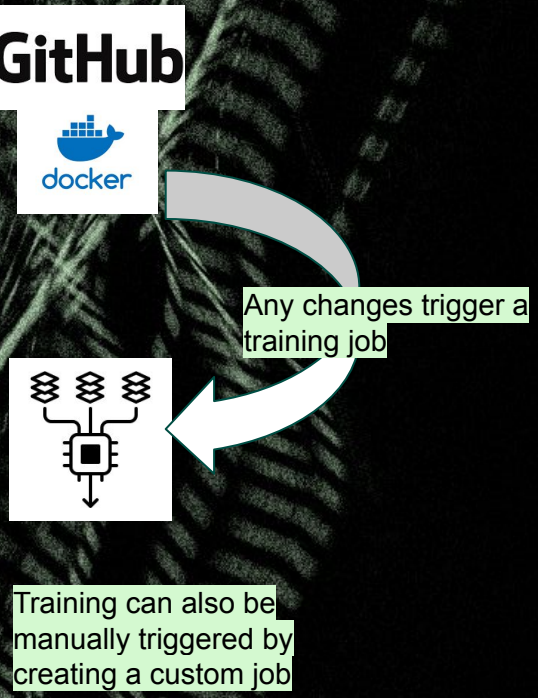
attention arc

Training Scripts

- → SDXL training script from huggingface diffusers github repo - https://github.com/huggingface/diffusers/blob/main/examples/text_to_image/train_text_to_image_sdxl.py
- → Cleaned up for easy readability and maintenance -
- → Breakdown into understandable smaller scripts, functions etc. - makes it easy for your future self and your team - “empathetic coding practices”

attention arc

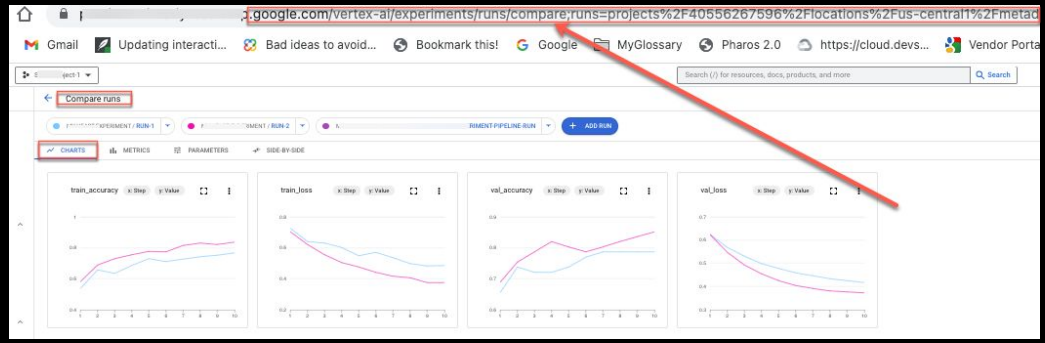
Training



- Containerize the training scripts - making it portable to any infrastructure (GCP, AWS, Azure etc.)
- Our choice of infrastructure for this project - GCP vertex AI
- Cloudbuild - automated deployment of the image - better tracking and saving time when updating the training image
- Custom job in vertex AI that uses the docker image
- AutoML for no-code train

attention arc

Tracking and Monitoring



Runs

Filter Enter property name or value

Name	Status	Type	Created
4f39024b-5[redacted]-a536-4[redacted]708a	Complete	Experiment run	Mar 14, 2025, 9:45:22 AM
e6c3775a-1[redacted]3-9cf9-ccdc077de49d	Complete	Experiment run	Mar 14, 2025, 9:43:23 AM
1fac8c38-a[redacted]-b3fd-c[redacted]fd1f	Complete	Experiment run	Mar 14, 2025, 9:40:32 AM

- → Vertex AI automatically tracks experiments
- → Can sort by accuracy or any other metric to retrieve the best model for inference

attention arc

Training in a dev environment

- → Compute Engine - https://docs.cloud.google.com/compute/docs/instances/create-start-instance#console
- → Virtual machine that supports any workload for development purposes
- → Similar to training locally, no containerization, no yaml files
- → Set up the compute engine, SSH into it from terminal or IDE of choice
- → MIflow or similar tools for experiment tracking

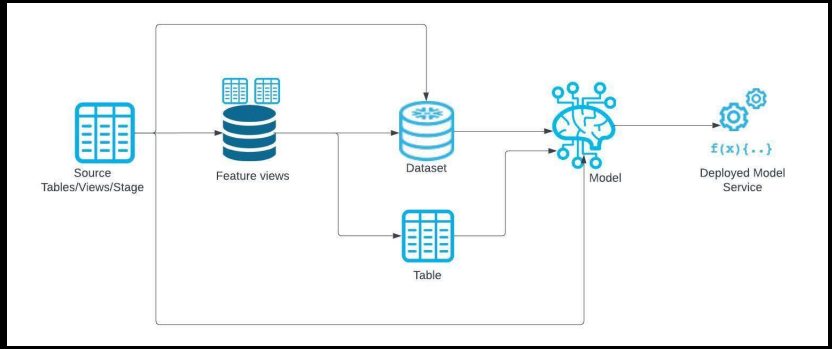
attention arc

Cost Breakdown

	Purchase Cost	Cloud cost per hour	Training hours
RTX 4090	$1599	$0.35	~4568
A100	$4699 (basic)	$4	~1175
H100	~$30000	$11	~2727

attention arc

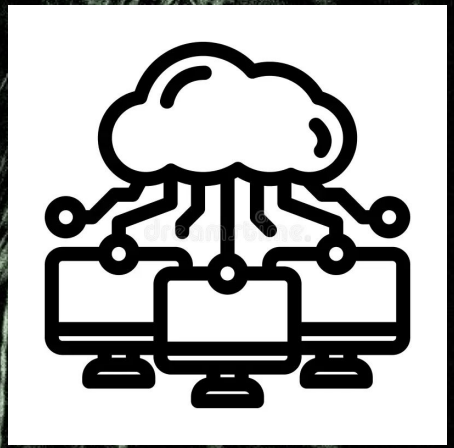
Snowflake ML



- → Save models to the model registry
- → Track experiments easily
- → Access to GPUs as well
- → Very useful if datasets are already in snowflake
- → Can save different versions with aliases for easy retrieval

attention arc

Advantages of Cloud Infrastructure



- → No need to own a GPU
- → Scale up and down with ease
- → Available to multiple developers
- → Model training uniformity for organizations
- → Cost effective for individuals and small organizations
- → Easier and reliable data storage and management
- → Tracking and monitoring for production systems
- → Data and processes hand-offs are seamless
- → Access to latest hardware instantly
- → Reduced infrastructure maintenance and costs
- → Faster onboarding for new developers

Environmental Impact

- Essentially sharing GPUs
- Maximizes GPU utilization and lifespan
- Reduces the number of GPUs manufactured and discarded.
- Cloud providers can reassign outdated versions for less demanding tasks
- They also have access to better energy efficiency
- Access to renewable energy at scale

attention arc

Find me here.

EMAIL

suha.mokalla@attentionarc.com

LINKEDIN (SCAN QR CODE →)

linkedin.com/in/suha-mokalla-ph-d-aa7612107/

FOLLOW ATTENTION ARC ON LINKEDIN

linkedin.com/company/attention-arc/

attention arc