

Pragalva Sapkota

Machine Learning | Language Models | ML Systems | Python

Kathmandu, Nepal | +977 9741860722 | pragalva.sapkota@deerwalk.edu.np
pragalva.me | github.com/PragalvaXFREZ | linkedin.com/in/pragalva-sapkota

PROJECTS

Embedding models from scratch

Sep 2026 – Present

Python, PyTorch, NumPy, Jupyter

[GitHub](#)

- Reimplementing word representations paper by paper: Bengio et al. (2003), word2vec, GloVe, and ELMo, continuing to the Transformer and BERT, each with an experiment and a write-up.
- Trained a 26.9M-parameter ELMo-style biLM on 101M words on an A100 (2.6 hours); pretrained representations lifted SST-2 sentiment accuracy from 62.0% to 78.7% and Penn Treebank POS accuracy from 78.2% to 96.5%.
- Replaced full softmax with hierarchical softmax in a skip-gram word2vec, cutting a projected 28-hour run over 11.46M training pairs to 350 seconds.
- Compared GloVe and word2vec across corpus sizes and seeds on WordSim-353 and analogies; showed GloVe's loss falling while its vectors got worse, traced to a weighting threshold only 0.09% of cells reached.

Edge inference benchmark, Jetson Nano

Sep 2026

TensorRT, ONNX Runtime, CUDA, NumPy, OpenCV, GStreamer

[GitHub](#) | [Write-up](#)

- Ran YOLOv8n through a TensorRT FP16 engine built on the board, with NumPy pre- and postprocessing verified against the Ultralytics reference on ONNX Runtime first.
- Timed every stage: 38.6 FPS for the network, 19.6 with pre/postprocessing, 12.3 live over RTP video; the GPU sat idle two thirds of each frame while four ARM cores did resizing, NMS, and video.

GPT from first principles

Jun 2026 – Present

Python, PyTorch, uv

[GitHub](#)

- Building a small GPT concept by concept: tokenization, embeddings, causal multi-head attention.

EXPERIENCE

R&D Department, DWIT

Sep 2026 – Present

Infrastructure Initiative Lead

Kathmandu, Nepal

- Selected to lead a student-run compute initiative; mentor three junior members, run a Proxmox VE host for disposable student environments, and proposed a GPU node for CUDA and model training to the department.

Linux Foundation Mentorship (LFX), Meshery (CNCF)

Mar 2026 – May 2026

LFX Mentee, Schema and API Governance; Meshery and Layer5 contributor Jan – Jun 2026

Remote

- Built a Go consumer-audit system covering 490+ API operations across three repositories; AST-based route parsers flag missing endpoints and type drift from schema contracts.
- Authored [100+ merged pull requests](#) and reviewed 180+ across Meshery and Layer5, spanning Go services, CLI end-to-end testing, CI/CD, and schema-driven code generation.

devata, bare-metal Kubernetes homelab

Dec 2025 – Present

Talos Linux, Argo CD, Cilium, Velero, Prometheus, Grafana

[GitHub](#) | [Live status](#)

- GitOps-managed Talos cluster (19 Argo CD Applications) with off-cluster backups and a live status page.

TECHNICAL SKILLS

Machine Learning: PyTorch, NumPy, TensorRT, ONNX Runtime, OpenCV, Jupyter

Languages: Python, Go, Bash, C

Infrastructure: Kubernetes, Docker, Linux, Proxmox VE, Argo CD, GitHub Actions, Prometheus, Grafana

WRITING

[ELMo: From One Word Vector to Context](#) (Oct 2026); [A Lower Loss Gave My GloVe Model Worse Vectors](#) (Oct 2026); [What My Word2Vec Model Learned](#) (Sep 2026); [Viewing Language Through a Probabilistic Lens](#) (Aug 2026).

EDUCATION

Tribhuvan University, Deerwalk Institute of Technology

Expected 2027

BSc in Computer Science and Information Technology (BSc CSIT)

Kathmandu, Nepal

LEADERSHIP

DeerHack 2025, Event Lead (Jan – Jul 2025): Led 50+ volunteers for 1,487 applicants; grew reach to 1M+.

DWIT Software Club, Vice President (Jul 2024 – Jul 2025): Ran club programs and the annual hackathon.