Deep Learning System Design

Table of Content

Introduction to Deep Learning System Design

- Defining Business and DL Objectives
- Requirements for Deep Learning Systems
- The Iterative Design Process
- Types of Deep Learning Tasks
- Classification vs. Regression
- Objective Functions
- Bridging the Gap Between Concept and Data

Data Acquisition and Preprocessing

- Sources of Data
- Third-Party Data Collection
- Data Serialization Formats Comparison
- Handling JSON, No-SQL, and Structured Data
- Transactional vs. Analytical Data Processing
- Extract, Transform, and Load (ETL) Processes
- REST and RESTful APIs

Training Data
- Sampling Techniques
- Labeling Strategies
- Addressing Class Imbalance
- Challenges with Imbalanced Data
- Modifying Loss Functions
- Data Augmentation Methods

Feature Engineering
- Handling Missing Values
- Scaling Features
- Data Discretization
- Encoding Categorical Features
- Feature Crossings
- Discrete and Continuous Positional Embeddings
- Preventing Data Leakage
- Engineering Effective Features

Model Architecture and Development
- Guidelines for Model Selection
- The Power of Ensembles (Boosting and Stacking)
- Experiment Tracking and Versioning
- Debugging Deep Learning Models
- Handling Model Failures
- Debugging Techniques
- Data and Model Parallelism
- AutoML (Soft and Hard)
- Four Phases of Model Development
- Offline Model Evaluation
- Evaluation Methods

Model Training and Evaluation

- Common Deployment Myths
- Batch vs. Online Prediction
- Unification of Batch and Streaming Pipelines
- Model Compression Techniques
- Utilizing Cloud and Edge for ML

Data Distribution Shifts and Monitoring

- Failures in Software Systems
- Failures Specific to ML
- Dealing with Edge Cases
- Detecting Data Distribution Shifts
- Monitoring and Observability

Continual Learning and Production Testing

- Concepts of Continual Learning
- Stateless vs. Stateful Retraining
- Challenges in Continual Learning
- Four Stages of Continual Learning
- Determining Model Update Frequency
- Testing Models in a Production Environment

Infrastructure and Tooling for DLOps

- Storage and Computational Resources
- Development Environments
- Cron, Schedulers, and Orchestrators
- Debugging and Maintenance Support
- Feature Stores for Efficient Data Management