Machine Learning Architecture & Design

Stages	Task	Goal
Data Collection System Design	system throughput, data storage, transfer, data format For example, capture images from sensor on-device	The collected data is identical to the data which the model will run on-device; scalability/efficiency
Data Collection	What and how much data to collect (categories, aspects), designing data collection protocol	Representative, training/val/test data,
Data Anotation/Clean	build ground truth (human labelers, third party algorithms) efficiently (quality and scalability) Handle missing data, outlier and noise data, unusable data	ground truth labels, evaluation dataset. Reasonable training dataset
Model Development	1. Architecture, loss function, optimization strategies (learning rate, batch size, others) 2. Addressing Overfitting/Underfitting/Model Collapse 3. Addressing regression cases 4. Model selection (simplicity, robustness, adversary attack)	The best model
Evaluation	Metric: accuracy, precision/recall considering unbalance cases Online/Offline evaluation Statistic/business metric	Extensive model perf eval
Model Deployment	1. Quantization(post-training quantization/quantization-aware training); 2. pruning (weight pruning/neuron pruning); 3.distillation; 4. customized to computer hardware; 5. memory management (dynamic memory allocation) and latency reduction (model partitioning, pipeline parallelism, asynchronous processing)	Computation, memory Latency
Analytic	1. Model performance metric monitoring 2. Model reliability and robustness: latency, throughput 3. Data/model drift/feature importance 4. User interaction/feedback: user satisfaction/engagement metric 5. Business metric: return on investment, conversion rates, others 6. Tool: model monitoring platform, custom dashboard, logging and alerting	Monitor performance for future improvement

The above stages outline a workflow for delivering machine learning-based feature. These stages are not necessarily sequential; they often overlap and are repeated.