Deep Learning Optimization

ai@electiconsulting.com

ELECTI

DEEP LEARNING OPTIMIZATION

Use Deep Learning efficiently at scale

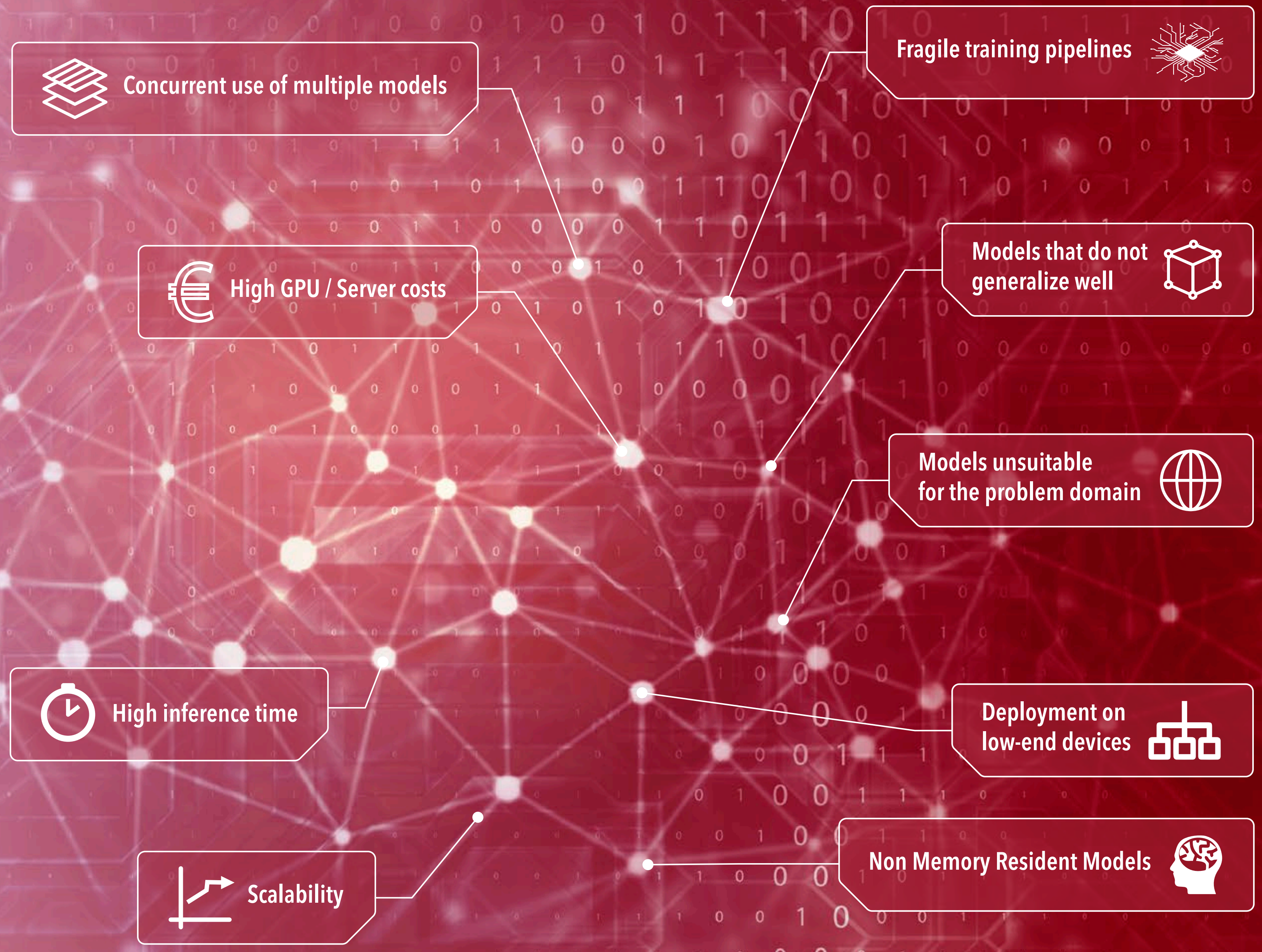
Deep neural networks have proved to be a very effective way to perform Machine Learning tasks. They excel when the input data is high-dimensional, the relationship between input and output is complicated, and the number of labeled training examples is large.

Nevertheless, top-performing ML systems can be expensive to store, slow to evaluate and hard to integrate into larger systems. We replace such cumbersome models with simpler ones that perform equally well, making them more efficient thus saving you time and resources.

Most companies employ off-the-shelf open source R&D models that are highly inefficient and very power hungry when deployed at scale. We work with GPU high performance environments such as TensorRT than can cut the inference time and resources needed by a huge margin thus saving server costs.

Services

Electi Consulting offers a variety of services to companies that have already gone the DL path but are facing challenges with:



Concurrent use of multiple models

High GPU / Server costs

High inference time

Scalability

Fragile training pipelines

Models that do not generalize well

Models unsuitable for the problem domain

Deployment on low-end devices

Non Memory Resident Models

Model Optimization

Our process for optimizing your DL pipeline follows these steps:

1- Model Architecture

2- Model Distillation

3-Multi-Task Optimization

4- Network Pruning

5- Network Quantization

6- Training Pipeline Optimization

7- GPU Optimization

1. Model Architecture

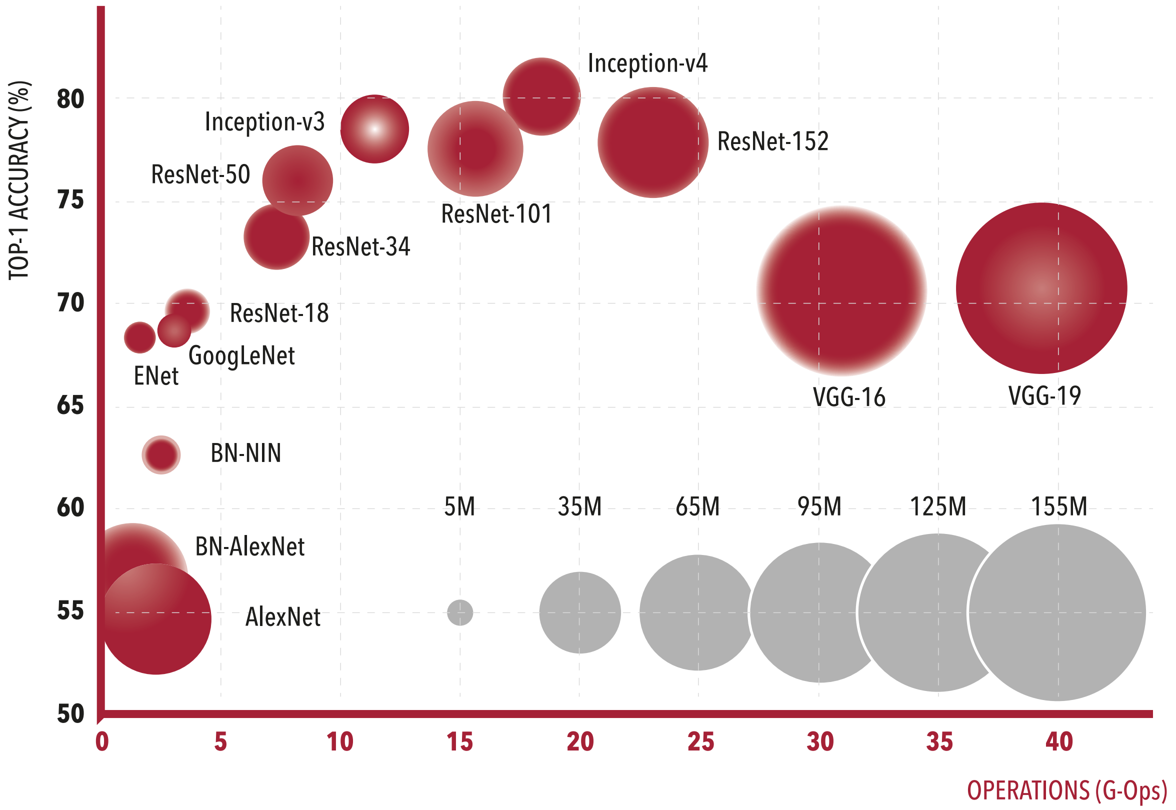
We analyze the problem you are solving, investigate the problem domain, analyze your current models and make suggestions based on state-of-the-art solutions and associated trade-offs involving:

PERFORMANCE

INFERENCE TIME

MEMORY FOOTPRINT

HARDWARE REQUIRED



In cases where the models used are not suitable for the problem at hand or are based on older generation assumptions we may suggest a re-modelling of the existing architecture.

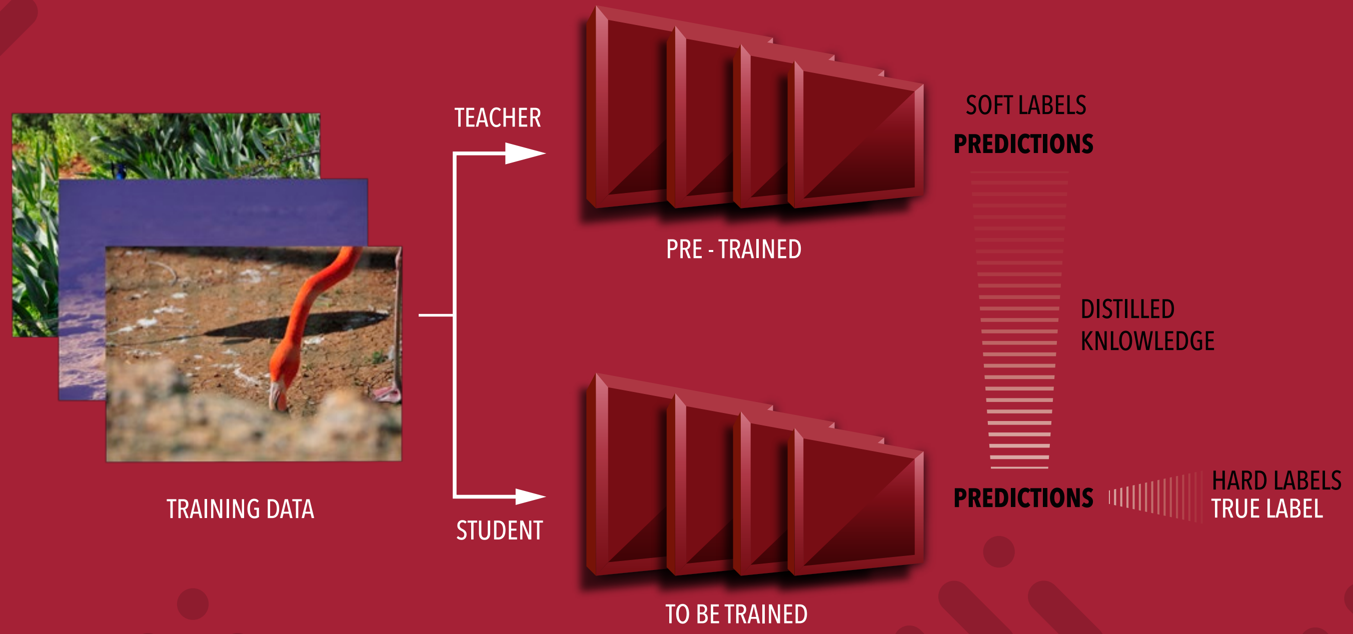
2· Model Distillation

Using the latest in model distillation research we convert your models into an equivalent smaller model while retaining minimal performance penalty. Depending on the domain we can achieve better performance than the original model.

Benefits:

LOWER INFERENCE TIME

SMALLER MEMORY FOOTPRINT



TEACHER

SOFT LABELS

PREDICTIONS

PRE - TRAINED

DISTILLED KNLOWLEDGE

TRAINING DATA

STUDENT

PREDICTIONS

HARD LABELS

TRUE LABEL

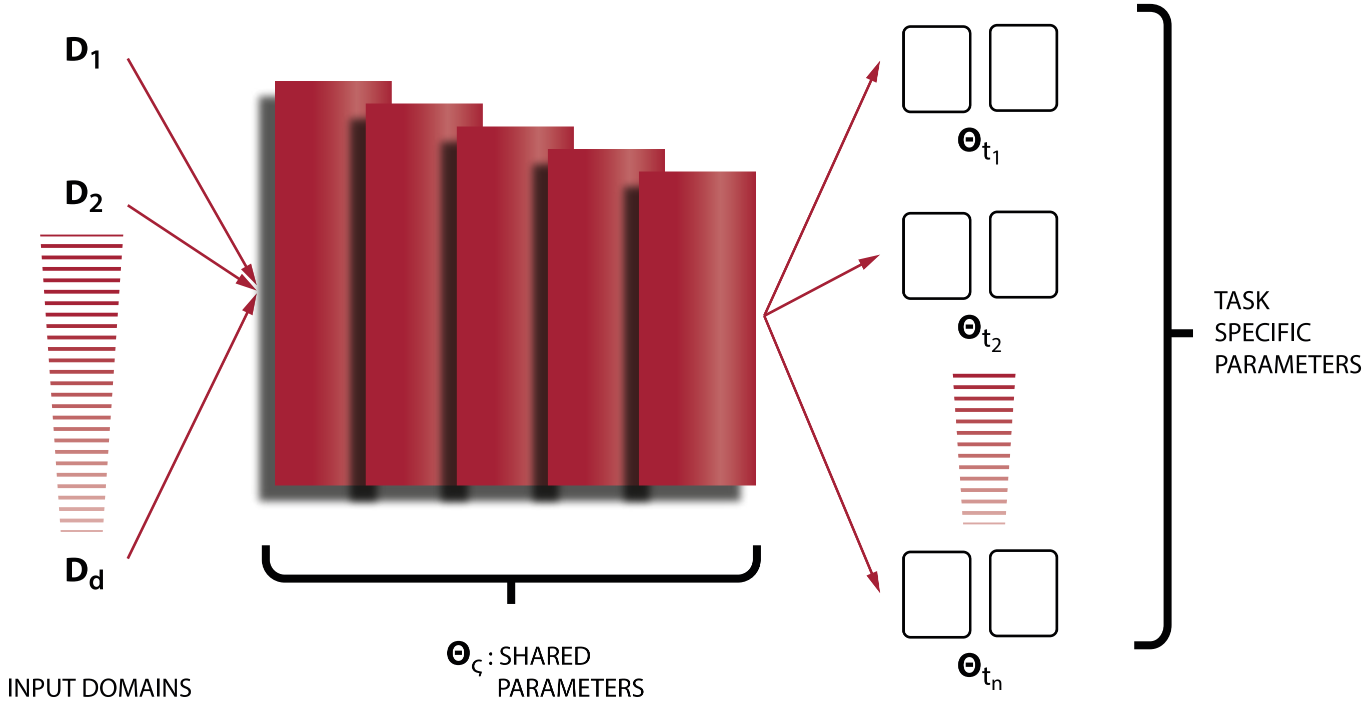
TO BE TRAINED

3· Multi-Task Optimization

If you are using multiple models we will investigate and suggest a model that performs multiple tasks at the same time. Such an approach usually generalizes better and has better inference time than the original models combined.

Benefits:

- LOWER INFERENCE TIME
- SMALLER MEMORY FOOTPRINT
- BETTER GENERALIZATION



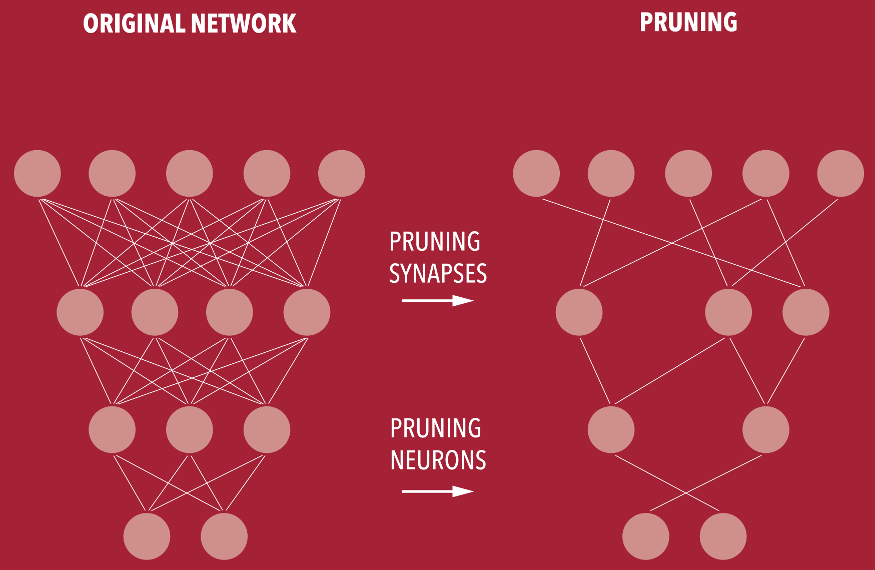
4· Network Pruning

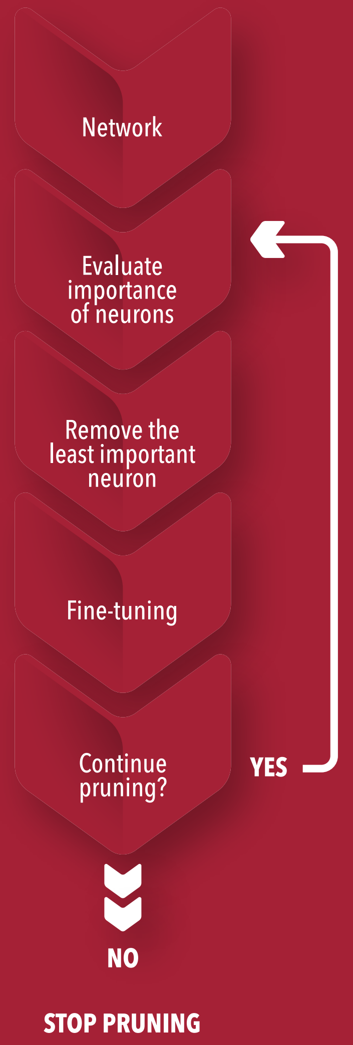
We retrain your system and remove redundant layers / channels whilst maintaining top performance. This method will keep the original architecture but remove excess neurons. We provide the tools for your team to integrate this new part in your pipeline.

Benefits:

LOWER INFERENCE TIME

SMALLER MEMORY FOOTPRINT




Network
Evaluate importance of neurons
Remove the least important neuron
Fine-tuning
Continue pruning?
YES
NO
STOP PRUNING

5· Network Quantization

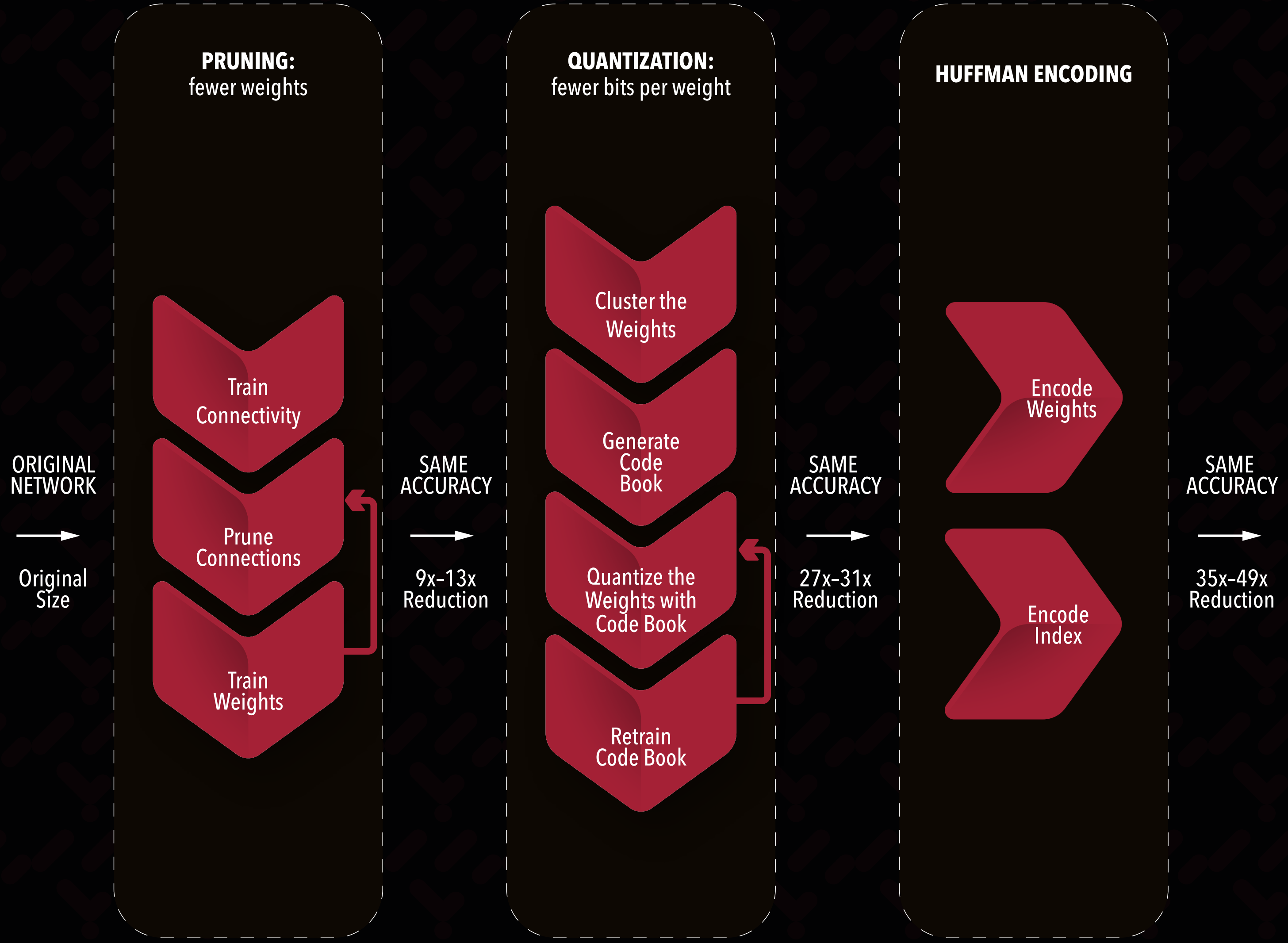
We quantize the model's weights thus saving memory and lowering the inference time. We achieve by methodically quantizing the network thus ensuring minimal drop in performance.

Benefits:

SAME ARCHITECTURE

LOWER INFERENCE TIME

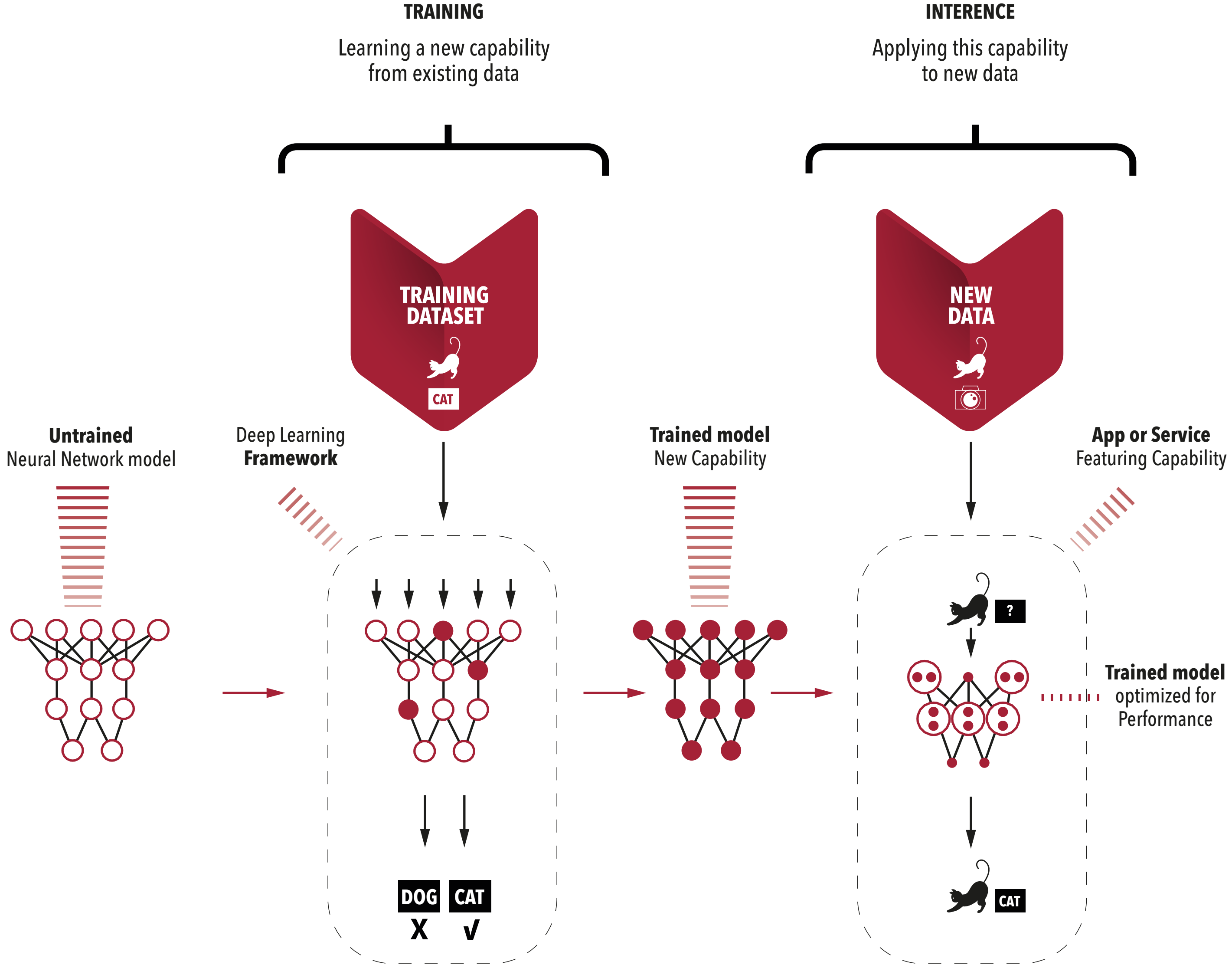
SMALLER MEMORY FOOTPRINT



6· Training Pipeline Optimization

We can help you setup or optimize your current DL training pipelines in several ways:

- Suggest alternative loss functions
- Identify data ingestion bottlenecks
- Add data augmentation steps to decrease required data samples
- Improve generalization and decrease required data samples by using Self-Supervised Learning methods



7·GPU Optimization

If you are in business serving millions of customers then you want to get the most out of your cloud or server **GPU**.

We can optimize and tune your solution for high performance and minimal cost.

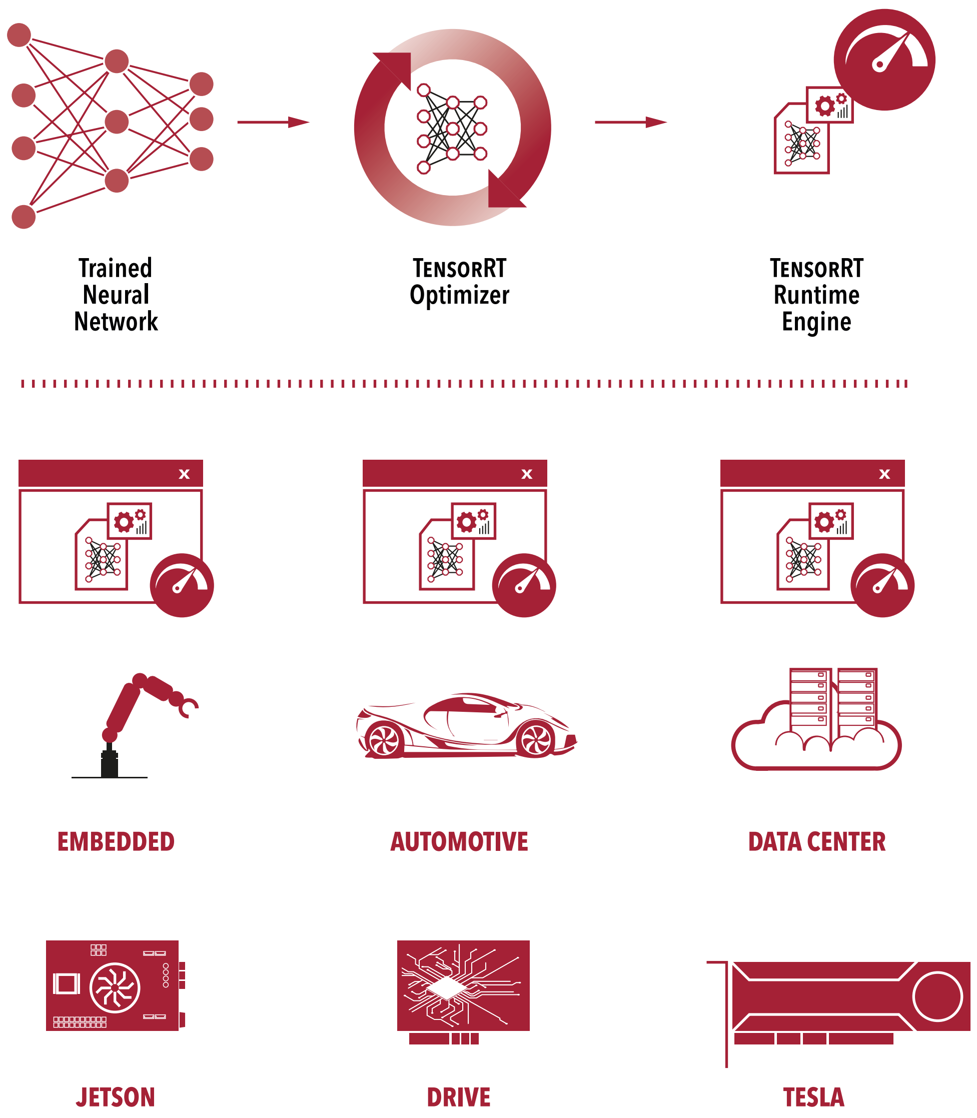
Specifically, our team can support you in migrating to the TensorRT framework which will result in a several-fold increase in **GPU** performance.

Benefits:

Optimisation with minimum cost

Increased GPU performance

Support



contact

ai@electiconsulting.com

visit

www.**electi**consulting.com