EuroMLSys ’25, Rotterdam, Netherlands

Decoupling Structural and Quantitative Knowledge in ReLU-based Deep Neural Networks

José Duato①②, Jose I. Mestre③, Manuel F. Dolz③, Enrique S. Quintana-Ortí④, José Cano⑤

①Qsimov Quantum Computing S.L., ②Royal Spanish Academy of Science
③Universitat Jaume I, Spain, ④Universitat Politècnica de València, Spain
⑤University of Glasgow, United Kingdom

Contents

Motivation: Cost and Complexity of DNN Training

Growing economic and environmental costs of DNN training.

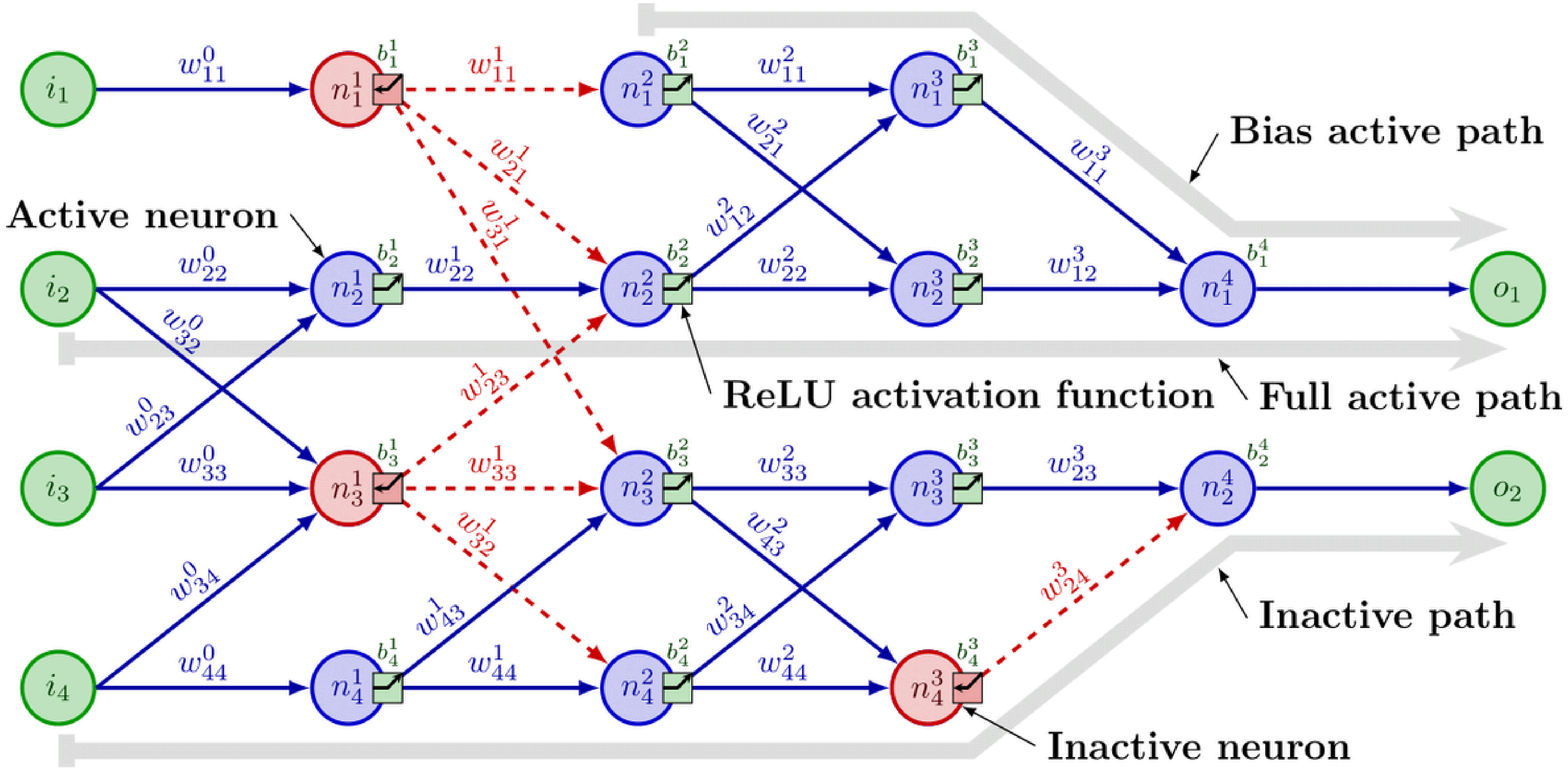
Retraining with evolving data is inefficient and costly.

Traditional DNNs couple linear and non-linear transformations.

Basic concepts

Active or Inactive Neuron: A neuron whose activation function (ReLU) produces a positive output for a given input; or a neuron that produces zero output.

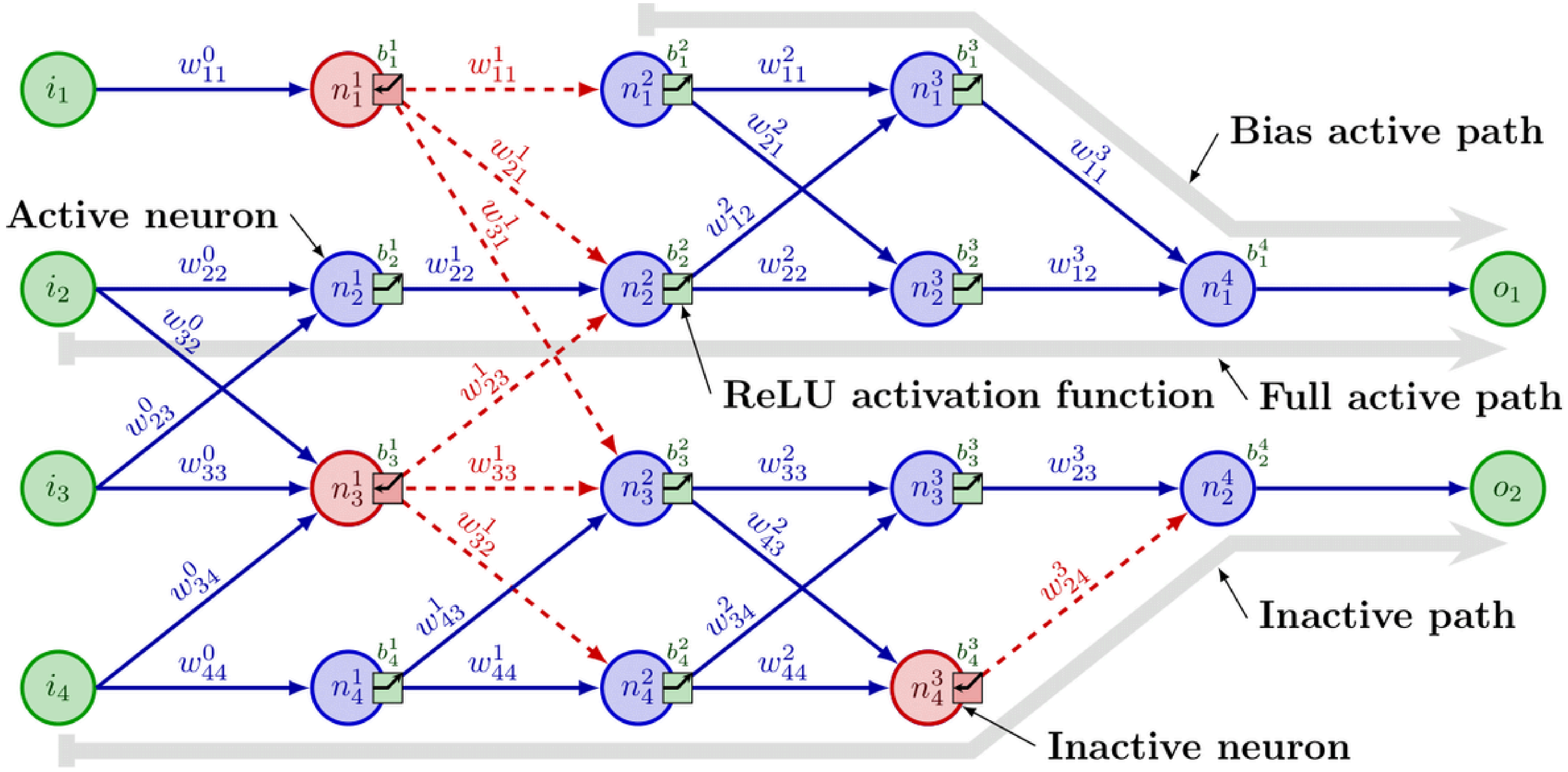
Activation Pattern: The specific set of active neurons corresponding to a given input.



Basic concepts

Activation Pattern: The specific set of active neurons corresponding to a given input.

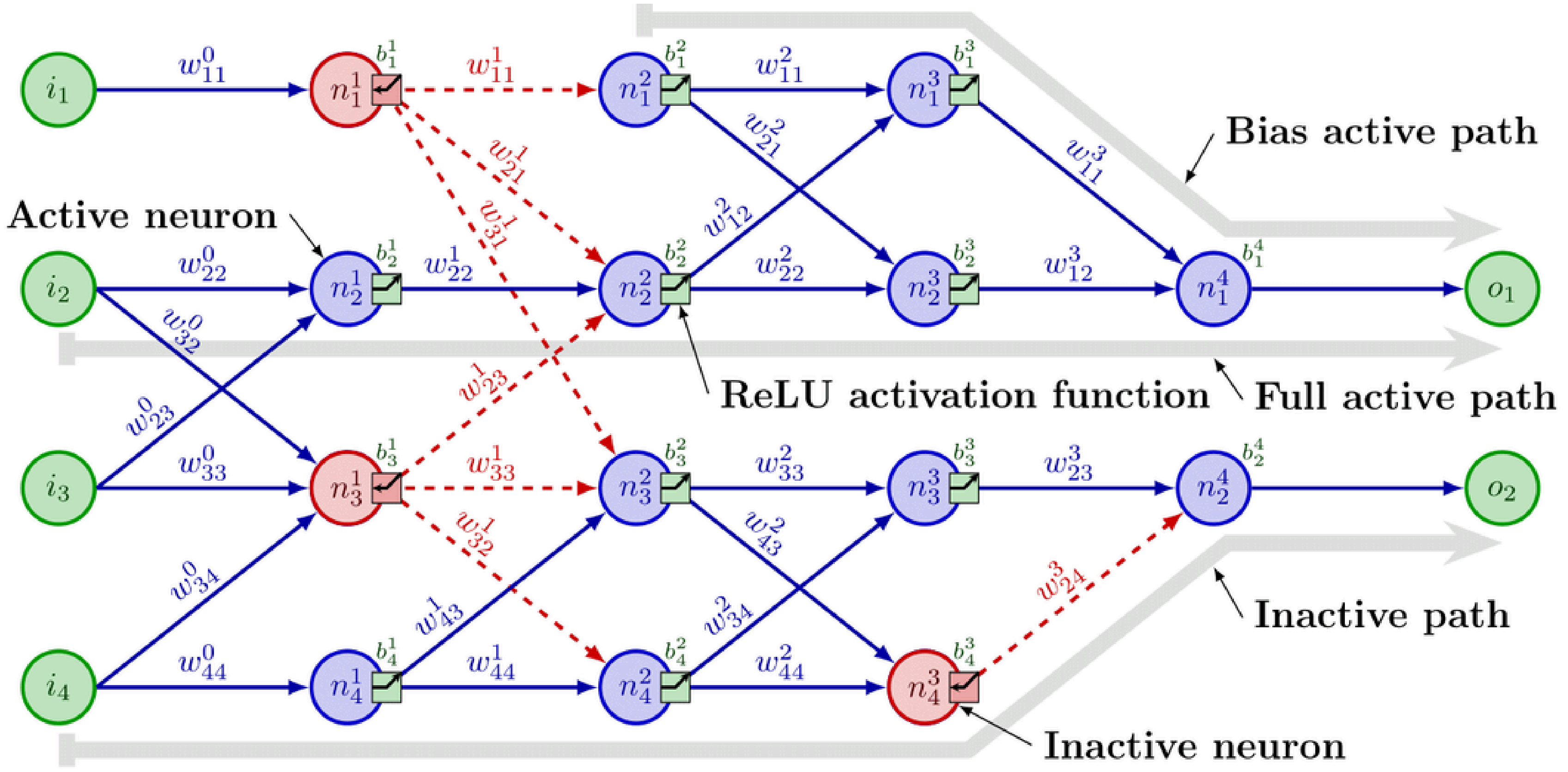
Active Path: A sequence of active neurons forming a continuous chain from input to output.



Basic concepts

Path Weight: The product of all the weights along a given path, including bias terms for bias paths.

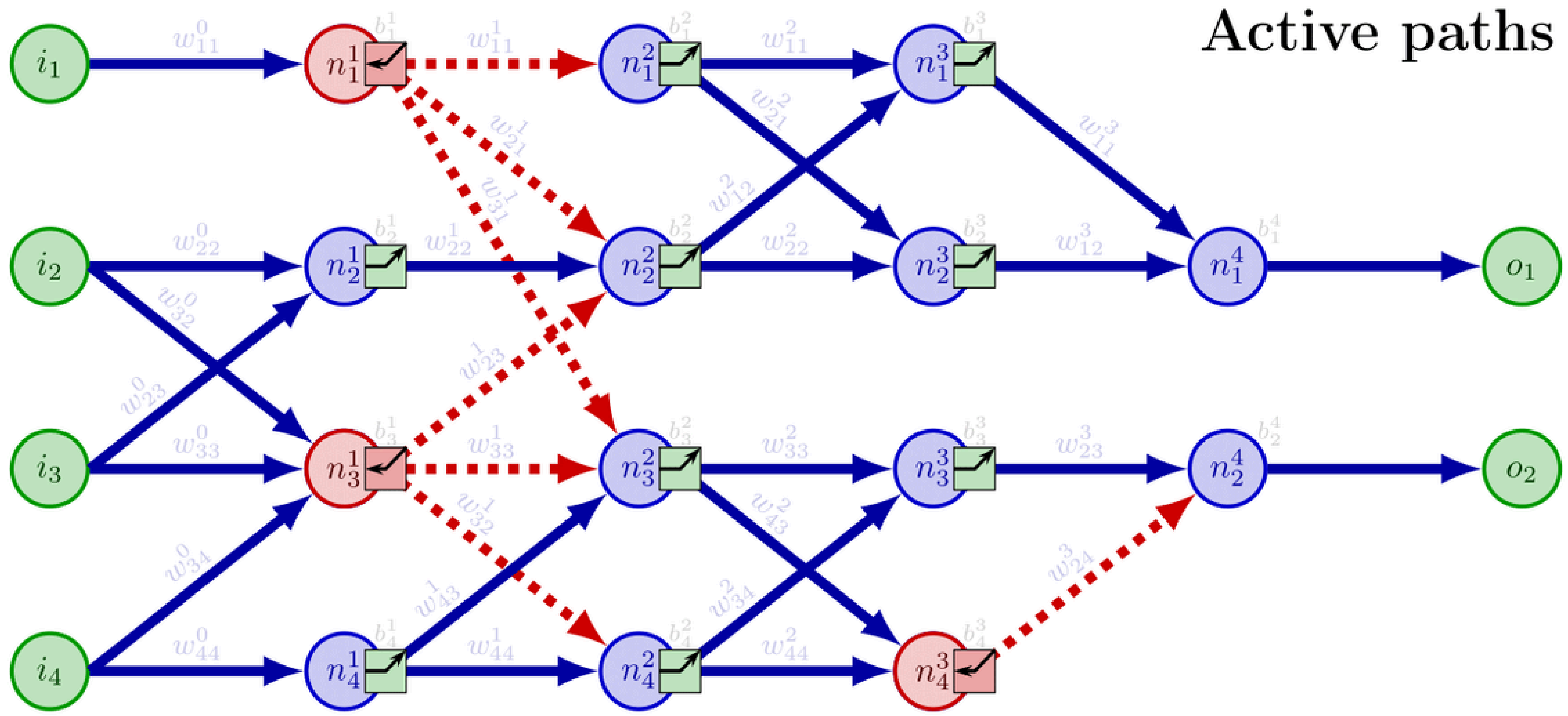
$$\begin{aligned} o_1 = b_1^4 \quad &+ \quad w^3_{11} b^3_1 + \boxed{w^3_{11} w^2_{11} b^2_1} + w^3_{11} w^2_{12} b^2_2 + w^3_{11} w^2_{12} w^1_{22} b^1_2 \\ &+ \quad \boxed{w^3_{11} w^2_{12} w^1_{22} w^0_{22}} i_2 + w^3_{11} w^2_{12} w^1_{22} w^0_{23} i_3 \\ &+ \quad w^3_{12} b^3_2 + \boxed{w^3_{12} w^2_{21} b^2_1} + w^3_{12} w^2_{22} b^2_2 + w^3_{12} w^2_{22} w^1_{22} b^1_2 \\ &+ \quad \boxed{w^3_{12} w^2_{22} w^1_{22} w^0_{22}} i_2 + w^3_{12} w^2_{22} w^1_{22} w^0_{23} i_3. \end{aligned}$$



Key Idea: Decoupling SK and QK in ReLU-based DNNs

Structural Knowledge:

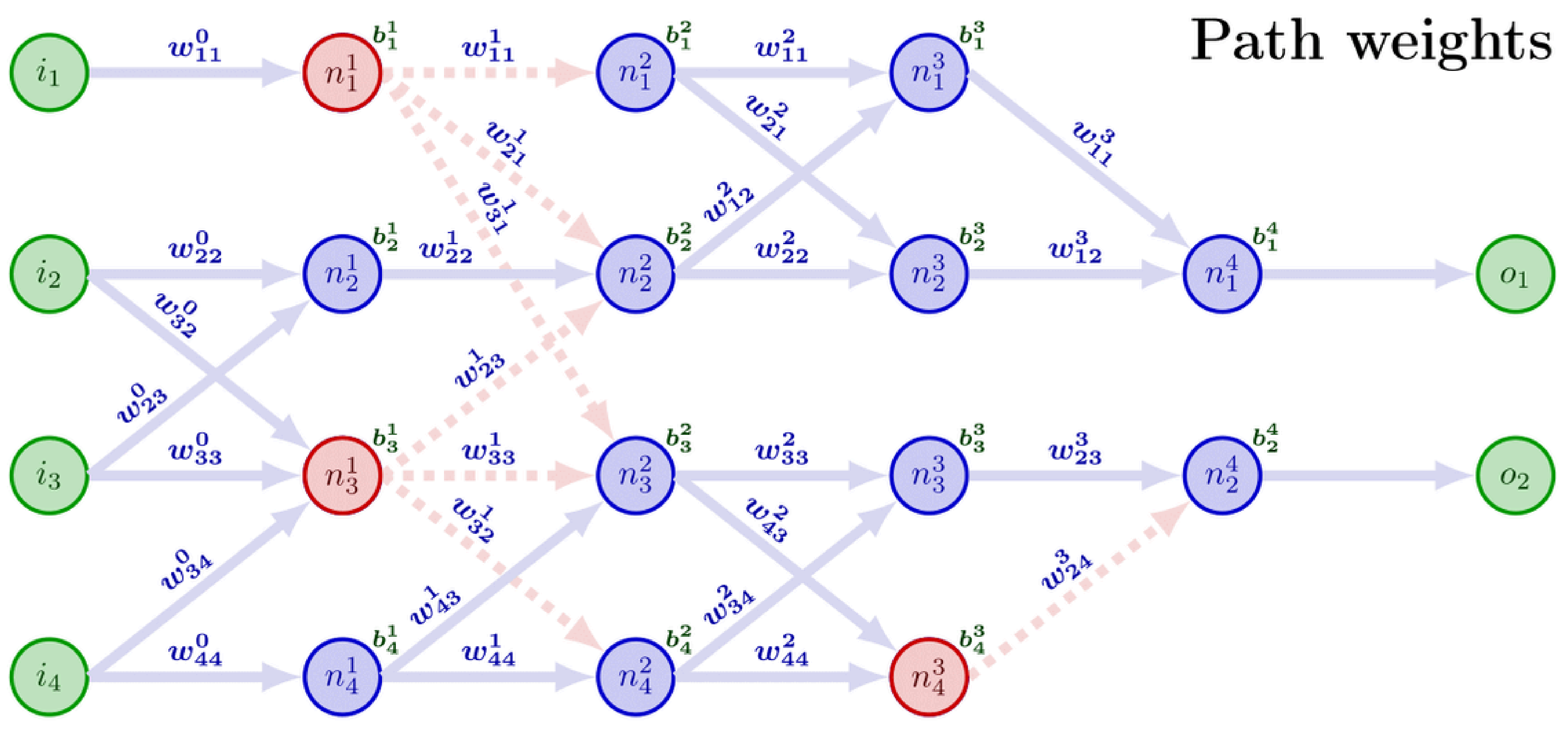
Determines which neurons and paths are activated for a given input, which captures the non-linear behaviour of the DNN.



Key Idea: Decoupling SK and QK in ReLU-based DNNs

Quantitative Knowledge:

Consists of the weights and biases used for computing outputs, turning the output calculation into a fully linear system.



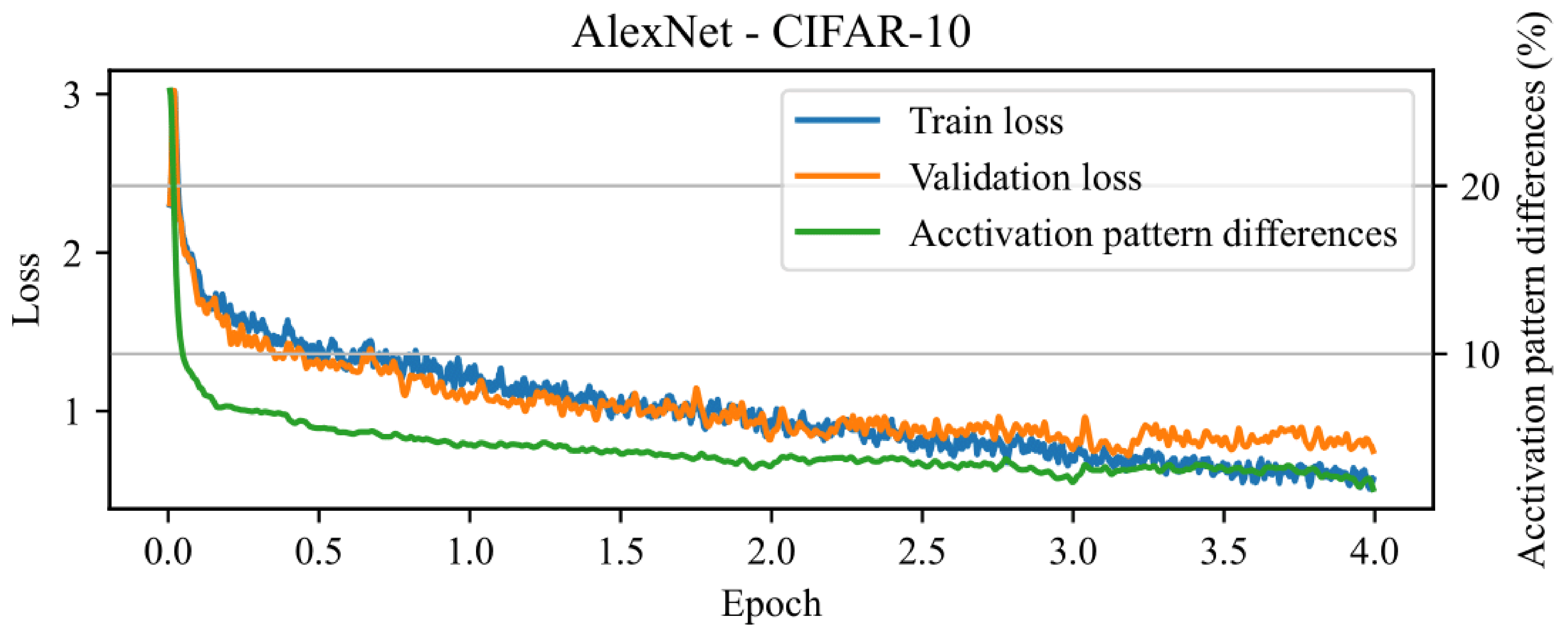
Key Idea: Decoupling SK and QK in ReLU-based DNNs

Structural Knowledge: Determines which neurons and paths are activated for a given input, which captures the non-linear behaviour of the DNN.
Quantitative Knowledge: Consists of the weights and biases used for computing outputs, turning the output calculation into a fully linear system.

Hypothesis 1: Structural Knowledge stabilizes quickly during training.
Hypothesis 2: Quantitative Knowledge can be re-trained and improve accuracy compared of training both Structural Knowledge and Quantitative Knowledge.

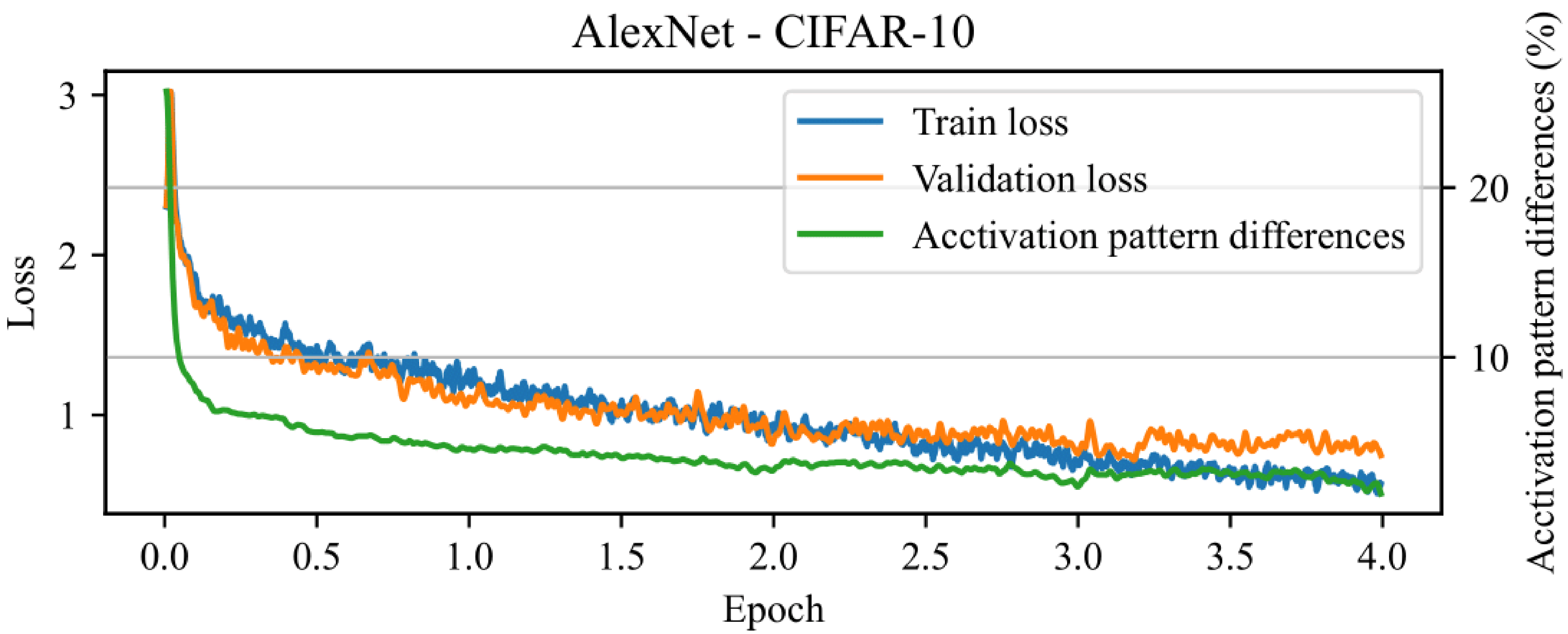
Experiment 1

Hypothesis 1: Structural Knowledge stabilizes quickly during training.



Experiment 1

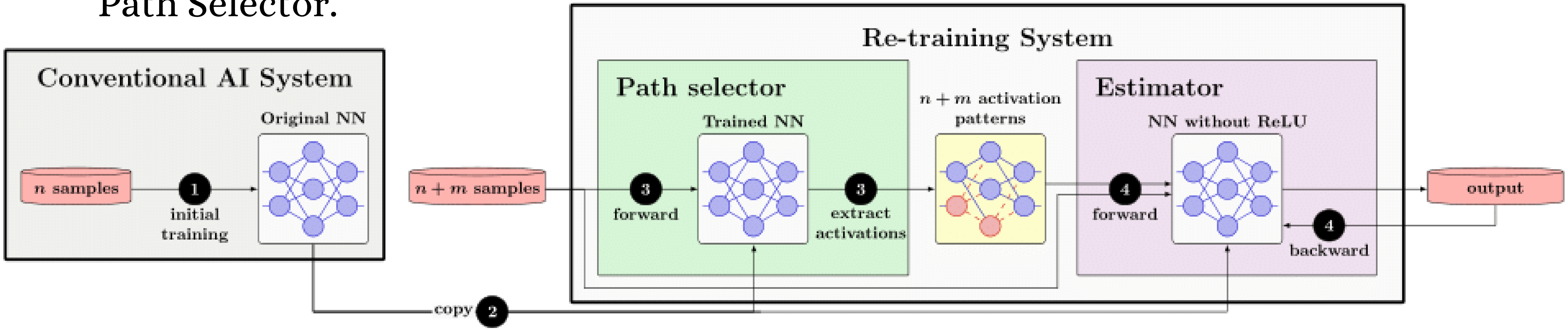
Activation pattern differences during training, measuring how often the activation state of the DNN neurons changes for a set of samples, compared to training and validation loss.



Proof-of-concept system

Path Selector: Extracts Structural Knowledge from a pre-trained model and is only responsible to determine wich paths are active.

Estimator: Trains only Quantitative Knowledge for improved efficiency, only the path weights are trained, while the path activation state is extracted by the Path Selector.



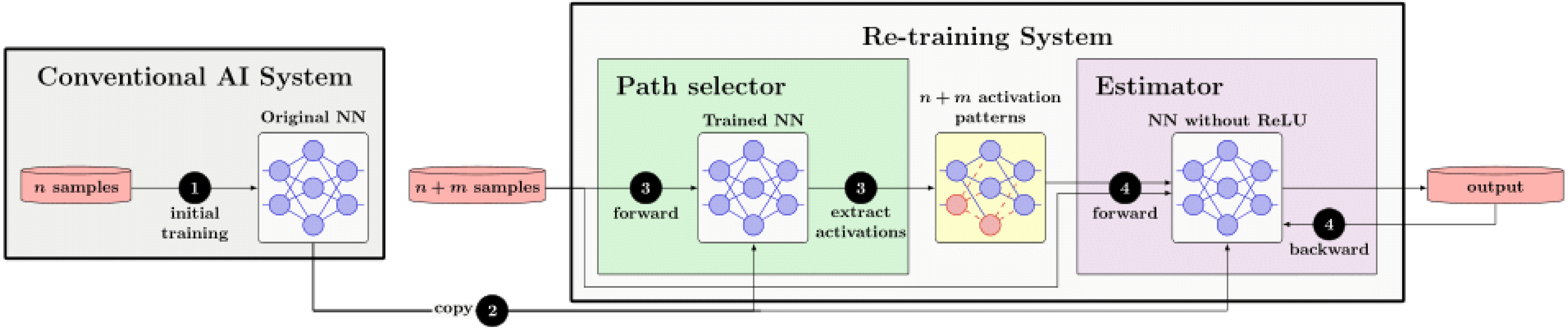
Experiment 2

Step 1: Initial training of DNN with n samples.

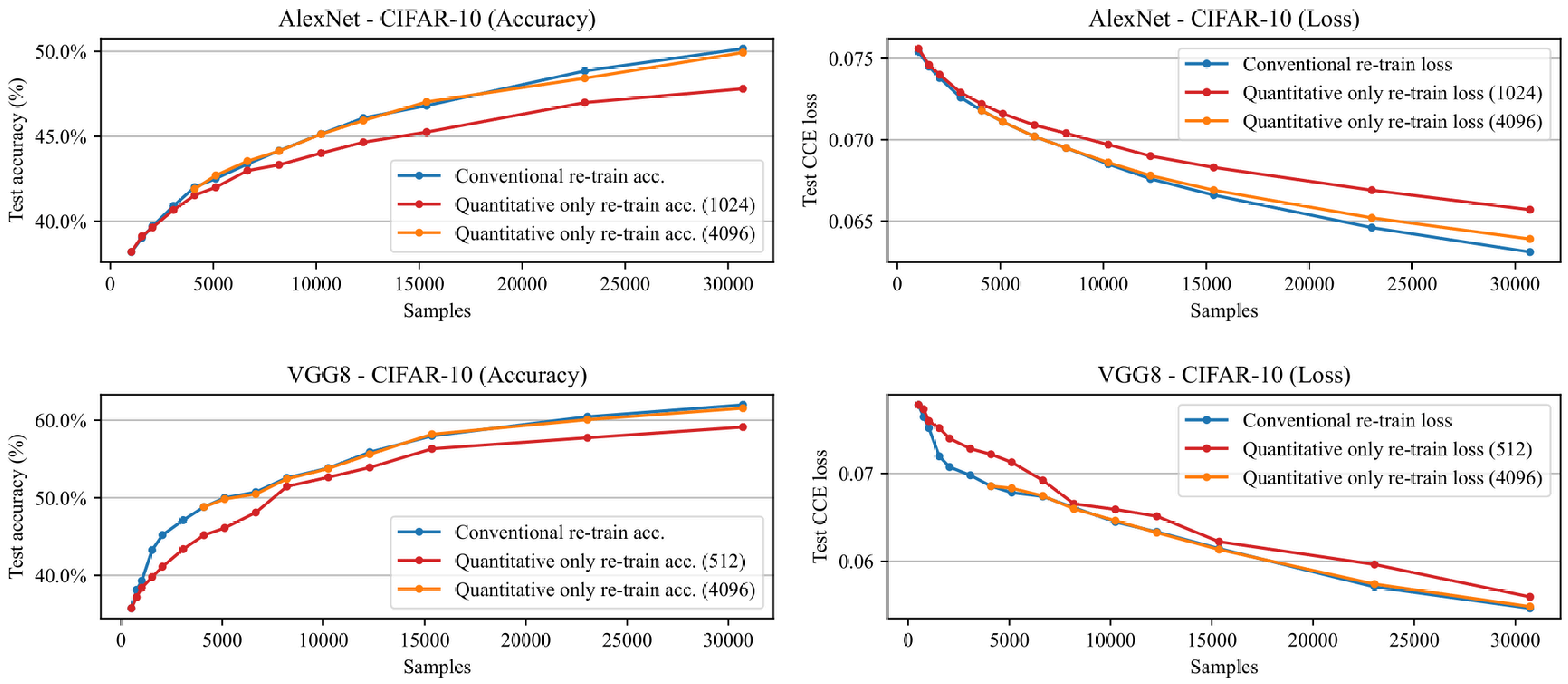
Step 2: One copy as Path Selector and one copy as initial Estimator.

Step 3: Inference with the Path Selector with n+m samples and extract activations.

Step 4: Train the Estimator with n+m samples.



Experiment 2



Conclusions

• The Structural Knowledge stabilizes faster than the Quantitative Knowledge during training.
• Decoupling the Structural Knowledge and the Quantitative Knowledge, freezing the Structural Knowledge and training only Quantitative Knowledge.

Future work: Develop an AI System that is capable of reducing training or retraining time, updating only the Quantitative Knowledge.