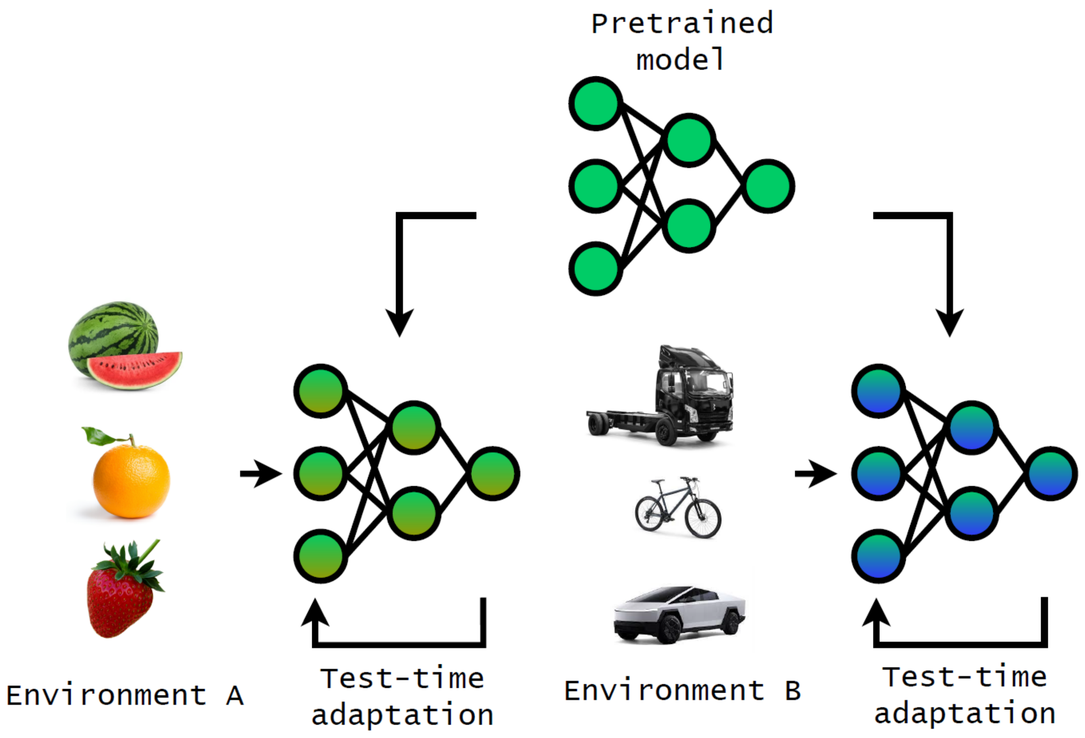
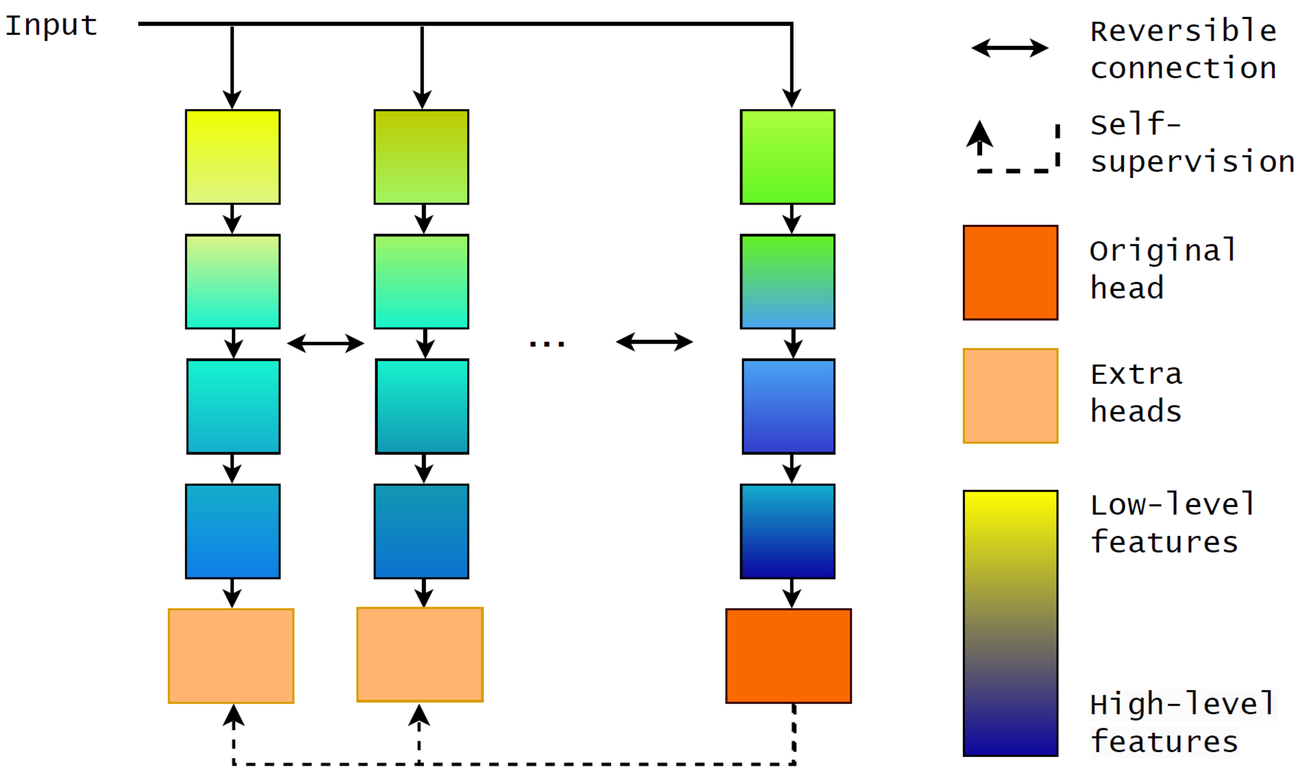
TEST-TIME SPECIALIZATION OF DYNAMIC NEURAL NETWORKS

Sam Leroux – Dewant Katare – Aaron Ding – Pieter Simoens

Introduction

- **Dynamic neural networks** can selectively use more or fewer computational resources at inference time based on the perceived complexity of the current input.
- We propose **an online specialization routine** that allows the model to **become more efficient** at recognizing classes that are observed frequently.
- The specialization procedure is efficient and **requires no human supervision.**
- The model remains capable of recognizing all classes, although samples of classes that are not observed frequently might require more calculations.



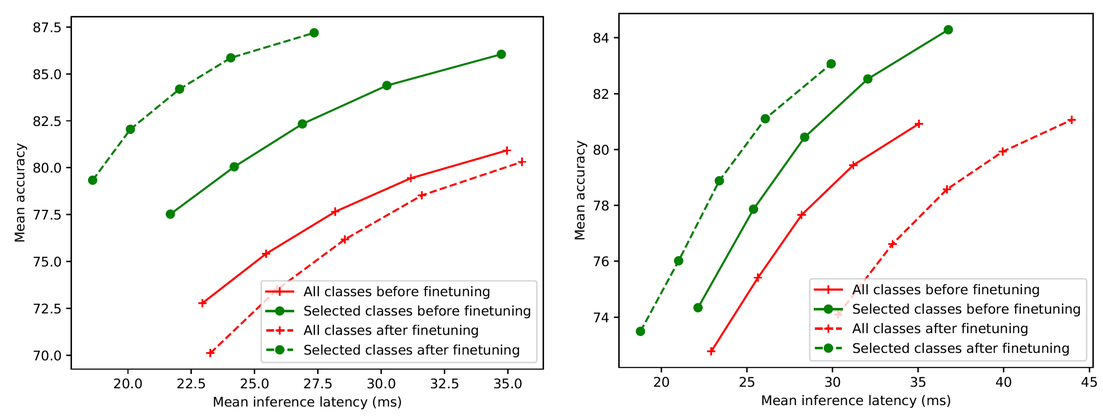


Approach

- We added multiple **early exits** to a RevCol model trained for image classification. Each exit (head) tries to make a prediction based on the features extracted so far. This prediction is only used if the confidence exceeds a predefined threshold.
- At inference time, we use the prediction of the selected exit as target to **update the earlier exits**, encouraging them to become better at recognizing similar objects in the future.
- We only update the head of the early, not-selected exits. **We do not update the backbone** to make sure that the external features remain compatible with all other layers of the model.
- Since only a very small part of the model is updated at inference time, the computational overhead is negligible.

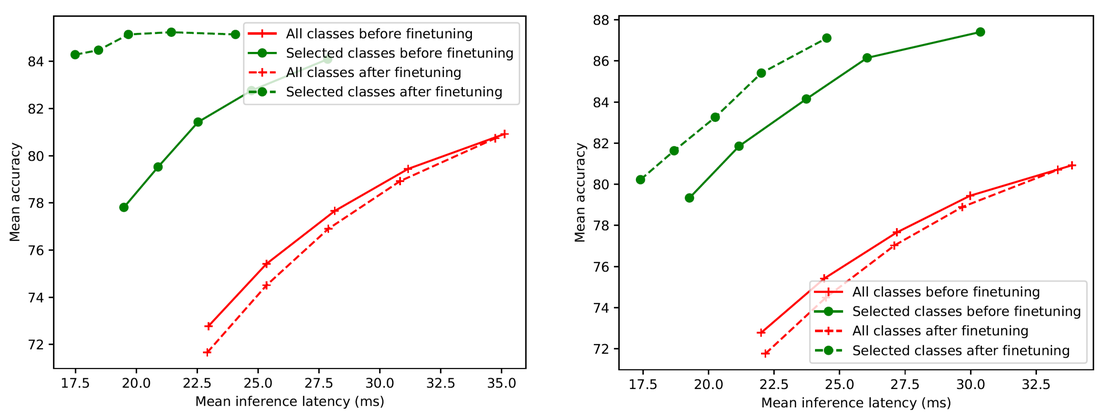
Results

- We trained the full model on the **ImageNet** dataset, obtaining a classification accuracy of 61.8%, 72.9%, 78.7% and 82.1% for each exit respectively.
- We then identified different subsets of ImageNet classes that contain similar types of objects (e.g., dogs, fruits, insects, cars, …).
- For each subset, we compare the performance of the model before and after specialization. By adjusting the threshold, we can select a trade-off between computational cost and accuracy. The solid curves on the right show this trade-off before specialization.
- The dashed curves illustrate how after specialization, the model becomes more efficient at recognizing the classes belonging to the subset while samples of other classes now require more calculations to achieve the same accuracy as before.
- All inference times were measured on an NVIDIA Jetson Orin development board.



Vehicles (42 classes)

Dogs (118 classes)



Fruits (21 classes)

Insects (27 classes)

Contact: Sam.Leroux@ugent.be

imec IDLab INTERNET & DATA LAB fwo TU Delft GHENT UNIVERSITY