Idea: propose a novel channel pruning approach that uses hardware behaviour to reshape networks

A small student network is trained both on the data and outputs of a larger pre-trained teacher network.

**Motivation**

- Inference time for a layer of ResNet-34 vs #channels on Intel Core i7
- **Staircase pattern:** For a given inference time, the green points maximise the network’s capacity (i.e. we get extra channels for no increase in time).

**Discovery and optimisation pipeline (1)**

- **Step 1:** Using channel saliency and empirical latency, design student

**Model distillation**

- Idea: propose a novel channel pruning approach that uses hardware behaviour to reshape networks
- A small student network is trained both on the data and outputs of a larger pre-trained teacher network.

**Discovery and optimisation pipeline (2)**

- **Step 2:** Train via attention transfer

**Results: CIFAR-10**

- **Student discovery algorithm**
  - Starting: base model, a Fisher-pruned reduction of the base model, and a target hardware platform
  - We iterate over all prunable layers in the base model and construct a set of optimal points
  - We then adapt the pruned layer widths to their nearest optimal point and return the resulting architecture

**Results: ImageNet**

- **Conclusions**
  - We have described a simple method for discovering performance enhanced reductions of baseline, large neural networks
  - We have compared our technique to common pruning approaches, and demonstrated its superiority on both the CIFAR-10 and ImageNet datasets for popular networks and hardware platforms