SimCLR v2

* This paper uses **unsupervised pre-training** and then **supervised fine-tuning**.
* It uses un-labelled data in a **task-agnostic** way.
* **Bigger networks** are used. Fewer the labels, the more this approach benefits from a bigger network.
* After fine-tuning, the big network can be further improved and distilled into a much smaller one with little loss in classification accuracy by using the unlabelled examples for a second time, but in a **task-specific way.**
* The algorithm can be summarized into three parts:
  + **Unsupervised pre-training** on a big model using simCLRv2,
  + **Supervised fine-tuning** on small labelled data,
  + **Distillation** with un-labelled data for refining and transferring the task-specific knowledge.
* **Unsupervised pre-training:**
  + Improved version of simCLR.
  + Larger ResNet models (deeper but less wide).
  + The largest model trained was ResNet152 with 3x wider channels and selective kernels, it is a channel-wise attention mechanism that improves the parameter efficiency.
  + The projection head was made deeper (3-layer deep instead of 2-layer deep).
  + Instead of throwing away the projection head after training, the model was fine-tuned from a middle layer (1st layer).
* **Supervised fine-tuning:**
  + Fine-tuning is done from the 1st layer of the MLP projection head.
  + Fine-tuning from the first layer of the MLP head is the same as adding a fully-connected layer to the base network and removing a fully-connected layer from the head, and the impact of this extra layer is contingent on the number of labelled examples during fine-tuning.
* **Distillation:**
  + The fine-tuned network was used as a teacher to impute labels for training a student network.
  + If the number of labelled examples is significant, we can combine distillation loss with ground truth labelled examples using a weighted combination.
  + This can be performed using students either with the same model architecture (which further improves the task-specific performance), or a smaller architecture (which leads to a compact model).