Knowledge Distillation

- training large image/text models is computationally expensive
- models often need up to 100 million parameters to achieve state-of-the-art performance
- training those models requires a lot of computational resources, e.g. GPUs, time and data
- for example, Meta's largest Llama 2 model was trained for more than 1.7 million GPU hours
  - used NVIDIA A100 80GB GPUs
  - if one would cost 5 USD per hour, the training would cost more than 8.5 million USD[1]
- infeasible to train such, or similar, models for researchers, students, or companies with limited resources
- transfer learning was first strategy to profit from large pretrained models by finetuning them on a specific task, for a, potential, different use case
- disadvantage is that model size does not change, finetuning is still computationally expensive, especially for large models
- other option is Knowledge Distillation (KD)
- here we do not finetune an existing model, but train a smaller model, the student model, to replicate, or rather predict, the outputs of a larger model, the teacher model, for a given sample
- can be applied for both supervised and self-supervised settings, meaning the teacher model has been trained in a supervised or self-supervised manner, respectively
  - former referred to as response-based KD, latter as feature-based KD
- has the advantage that the student model can be much smaller than the teacher model, and can, depending which of the former settings is used, have a different architecture
- emirically shown that student models much smaller than teacher models can achieve similar performance

Response-based Knowledge Distillation

- teacher model is/was trained in a supervised manner
- provides logits as predictions for a given sample
- are the target of the student model
- logits are regressed by output of student model, which usually has the same shape
  - in that case mean squared error (MSE) is used as loss function
- student model can also regress the probabilty distribution of the teacher model, so the output of the teacher after softmax has been applied on logits
  - in that case Kullback-Leibler divergence (KLD) is used as loss function

Feature-based Knowledge Distillation

- teacher model is/was trained in a self-supervised, or supervised, manner
- student model tries to replicate/predict the (intermediate) activations of the teacher model
  - so not necessarily only the output of the teacher model, although this is also possible
- if teacher model has been trained in a self-supervised manner, feature-based is needed, as the teacher model does not provide logits or a probabilty distribution to regress
- if teacher model has been trained in a supervised manner, feature-based can also be used, but response-based is more common
- here usually MSE is used as loss function

[1] Calculation done based on the price per GPU hour of the NVIDIA A100 80GB GPU on AWS for instance p4de.24xlarge, as of July 2024.