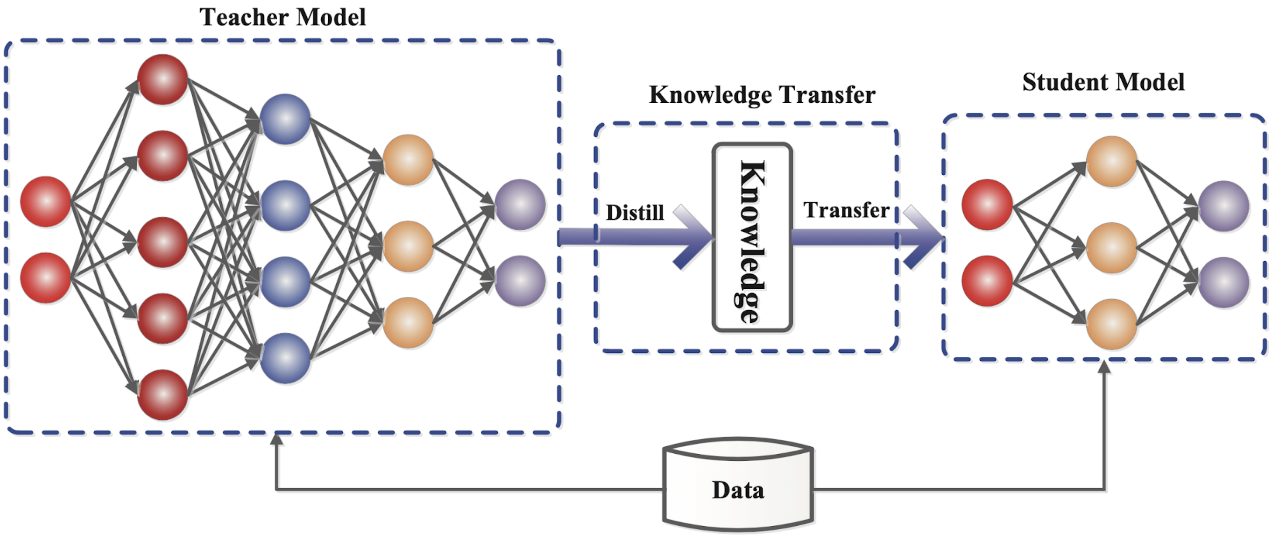
Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

Knowledge-Distillation

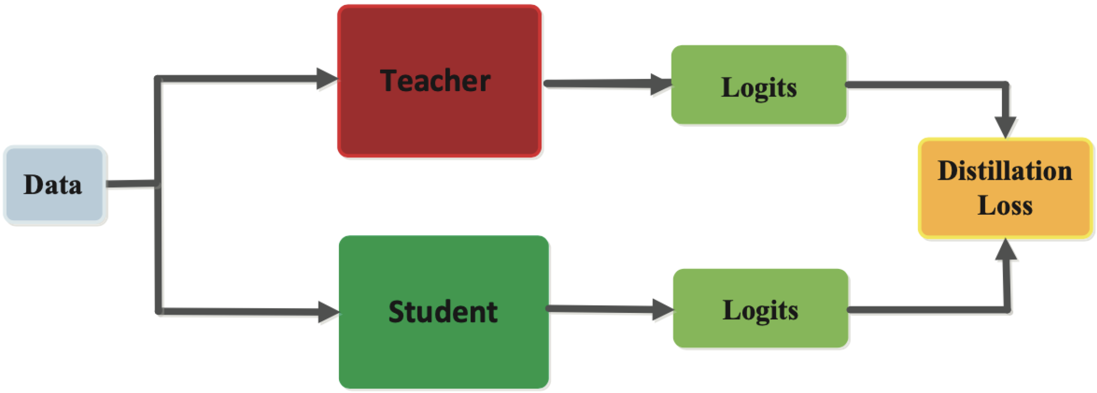
Knowledge distillation is a machine learning technique that aims to transfer the learnings of a large pre-trained model, the “teacher model,” to a smaller “student model.” It’s used in deep learning as a form of model compression and knowledge transfer, particularly for massive deep neural networks.



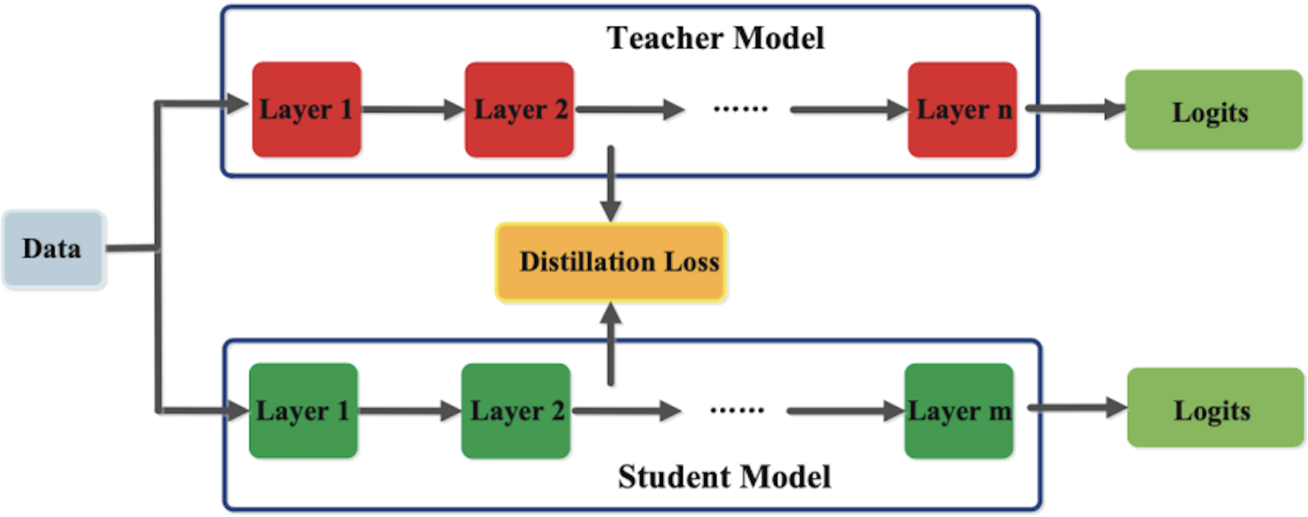
The goal of knowledge distillation is to train a more compact model to mimic a larger, more complex model. Whereas the objective in conventional deep learning is to train an artificial neural network to bring its predictions closer to the output examples provided in a training data set, the primary objective in distilling knowledge is to train the student network to match the predictions made by the teacher network.

Knowledge distillation (KD) is most often applied to large deep neural networks with many layers and learnable parameters. This process makes it particularly relevant to the ongoing proliferation of massive generative AI models with billions of parameters.

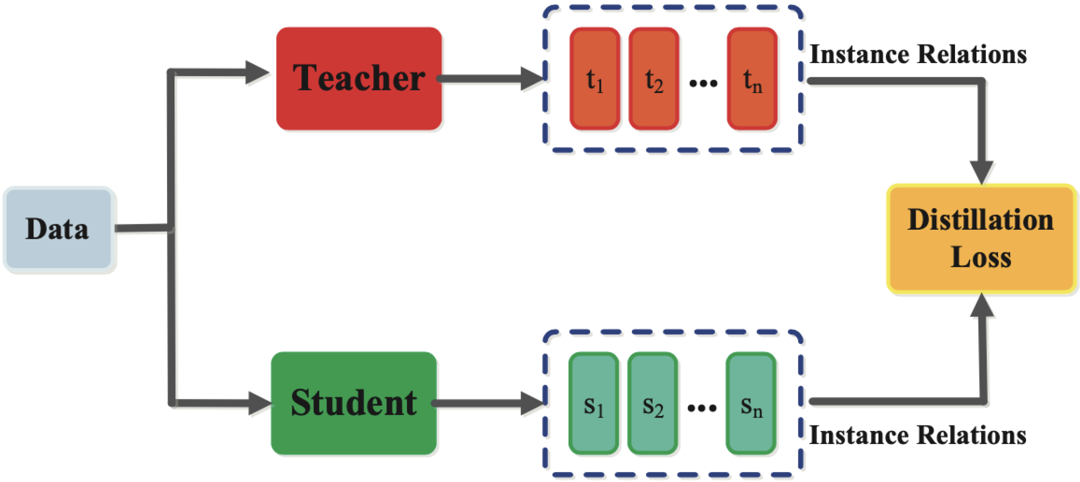
1. **Response-based Knowledge Distillation**



2. **Feature-based Knowledge Distillation**



3. Relation-based Knowledge Distillation



Why is Knowledge Distillation Important?

In many real-world settings, an artificial intelligence model's accuracy and capacity are not, unto themselves, enough to make the model useful: it must also fit within the available budget of time, memory, money and computational resources.

The top performing models for a given task are often too large, slow or expensive for most practical use cases—but often have unique qualities that emerge from a combination of their size and their capacity for pre-training on a massive quantity of training data. These emergent abilities are especially evident in autoregressive language models, like GPT or Llama, that exhibit capabilities beyond their explicit training objective of simply predicting the next word in a sequence. Conversely, small models are faster and less computationally demanding, but lack the accuracy, refinement and knowledge capacity of a large model with far more parameters.
