Knowledge Distillation: Principles, Algorithms, Applications

https://neptune.ai/blog/knowledge-distillation

Large-scale machine learning and deep learning models are increasingly common. For instance, GPT-3 is trained on 570 GB of text and consists of 175 billion parameters. However, whilst training large models helps improve state-of-the-art performance, deploying such cumbersome models especially on edge devices is not straightforward.

Additionally, the majority of data science modeling work focuses on training a single large model or an ensemble of different models to perform well on a hold-out validation set which is often not representative of the real-world data.

This discord between training and test objectives leads to the development of machine learning models that yield good accuracy on curated validation datasets but often fail to meet performance, latency, and throughput benchmarks at the time of inference on real-world test data.

Knowledge distillation [1] helps overcome these challenges by capturing and "distilling" the knowledge in a complex machine learning model or an ensemble of models into a smaller single model that is much easier to deploy without significant loss in performance.

In this blog, I will:

- describe knowledge distillation in detail, its underlying **principle, training schemes, and algorithms**;
- dive deeper into **applications** of Knowledge Distillation in deep learning for images, text, and audio.

What is knowledge distillation?

Knowledge distillation refers to the process of transferring the knowledge from a large unwieldy model or set of models to a single smaller model that can be practically deployed under real-world constraints. Essentially, it is a form of model

compression that was first successfully demonstrated by Bucilua and collaborators in 2006 [2].

Knowledge distillation is performed more commonly on neural network models associated with complex architectures including several layers and model parameters. Therefore, with the advent of deep learning in the last decade, and its success in diverse fields including speech recognition, image recognition, and natural language processing, knowledge distillation techniques have gained prominence for practical real-world applications [3].

The challenge of deploying large deep neural network models is especially pertinent for edge devices with limited memory and computational capacity. To tackle this challenge, a model compression method was first proposed [2] to transfer the knowledge from a large model into training a smaller model without any significant loss in performance. This process of learning a small model from a larger model was formalized as a “Knowledge Distillation” framework by Hinton and colleagues [1].

As shown in Figure 1, in knowledge distillation, a small “student” model learns to mimic a large “teacher” model and leverage the knowledge of the teacher to obtain similar or higher accuracy. In the next section, I will delve deeper into the knowledge distillation framework and its underlying architecture and mechanisms.

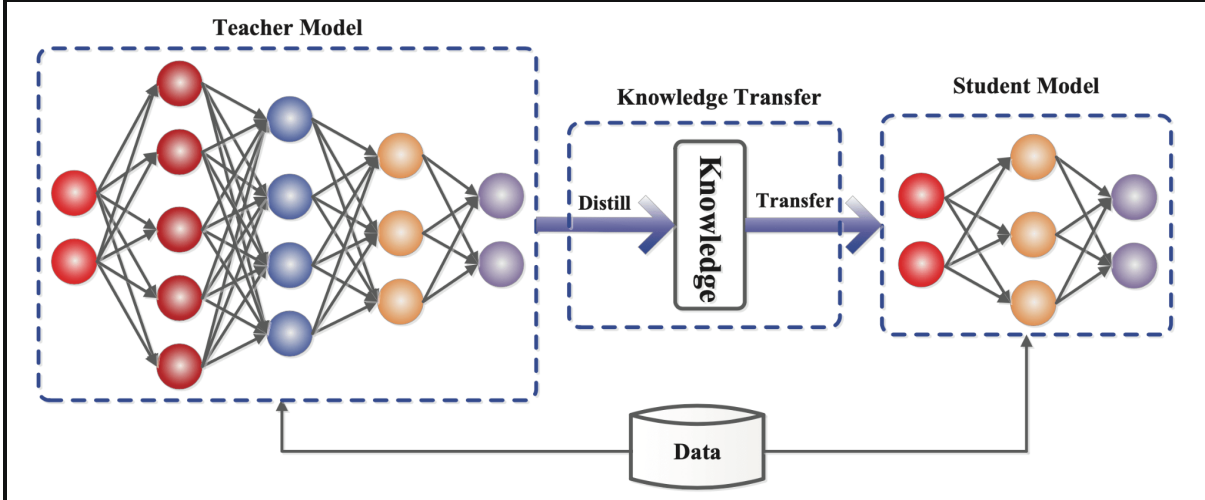

Figure 1. The teacher-student framework for knowledge distillation | Source: Arxiv

Diving deeper into knowledge distillation

A knowledge distillation system consists of three principal components: the knowledge, the distillation algorithm, and the teacher-student architecture [3].

Knowledge

In a neural network, knowledge typically refers to the learned weights and biases. At the same time, there is a rich diversity in the sources of knowledge in a large deep neural network. Typical knowledge distillation uses the logits as the source of teacher knowledge, whilst others focus on the weights or activations of intermediate layers. Other kinds of relevant knowledge include the relationship between different types of activations and neurons or the parameters of the teacher model themselves.

The different forms of knowledge are categorized into three different types: **Response-based knowledge**, **Feature-based knowledge**, and **Relation-based knowledge**. Figure 2 illustrates these three different types of knowledge from the teacher model. I will discuss each of these different knowledge sources in detail in the following section.

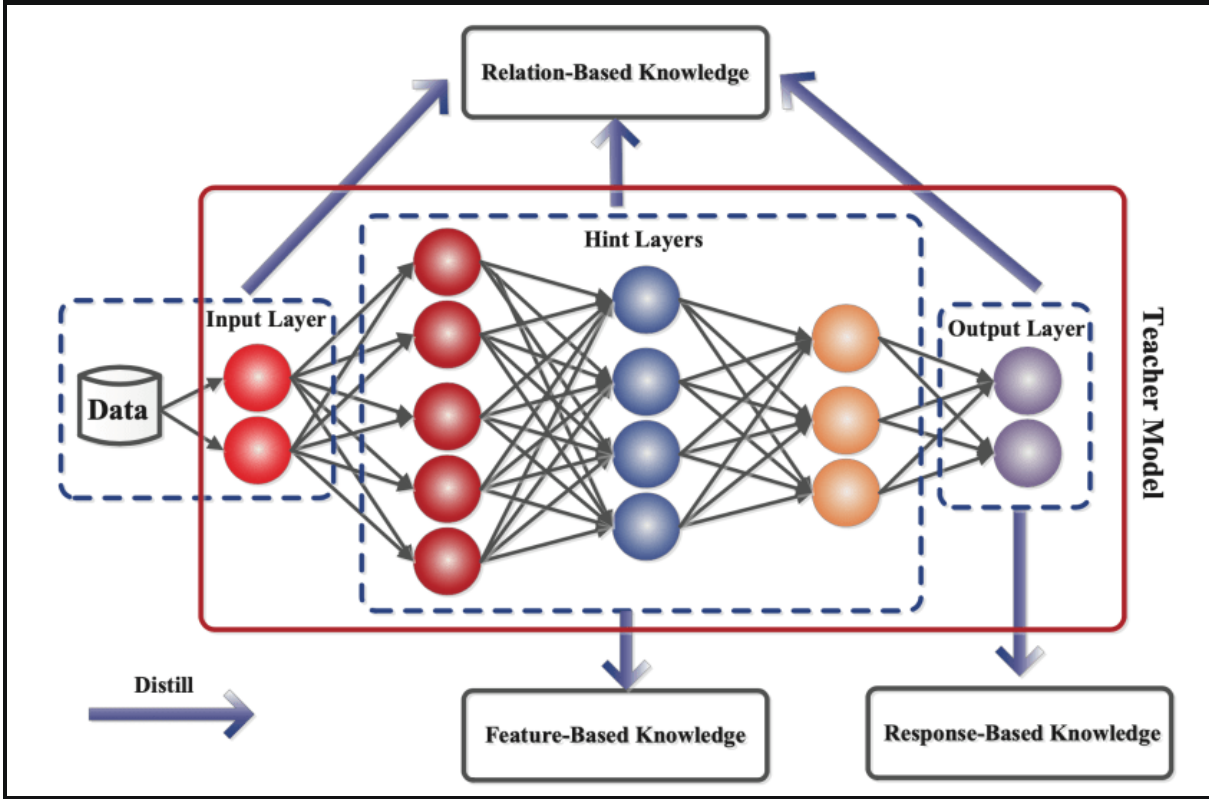

Figure 2. The different kinds of knowledge in a teacher model | Source: Arxiv

1. Response-based knowledge

As shown in Figure 2, response-based knowledge focuses on the final output layer of the teacher model. The hypothesis is that the student model will learn to mimic the predictions of the teacher model. As illustrated in Figure 3, This can be achieved by using a loss function, termed the distillation loss, that captures the difference between the logits of the student and the teacher model respectively. As this loss is minimized over training, the student model will become better at making the same predictions as the teacher.

In the context of computer vision tasks like image classification, the soft targets comprise the response-based knowledge. Soft targets represent the probability distribution over the output classes and typically estimated using a softmax function. Each soft target's contribution to the knowledge is modulated using a parameter called temperature. Response-based knowledge distillation based on soft targets is usually used in the context of supervised learning.

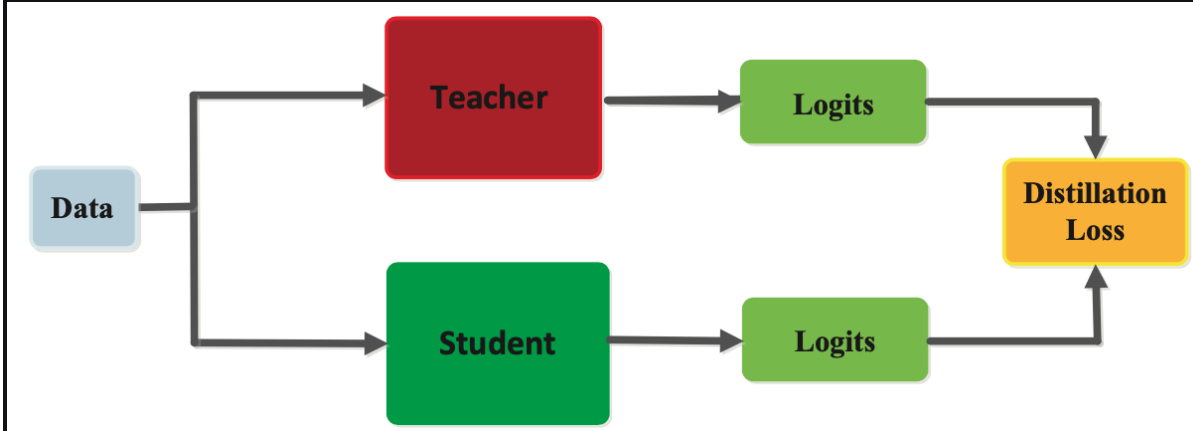

Figure 3. Response-based knowledge distillation | Source: Arxiv

2. Feature-based knowledge

A trained teacher model also captures knowledge of the data in its intermediate layers, which is especially pertinent for deep neural networks. The intermediate layers learn to discriminate specific features and this knowledge can be used to train a student model. As shown in Figure 4, the goal is to train the student model to learn the same feature activations as the teacher model. The distillation loss function achieves this by minimizing the difference between the feature activations of the teacher and the student models.

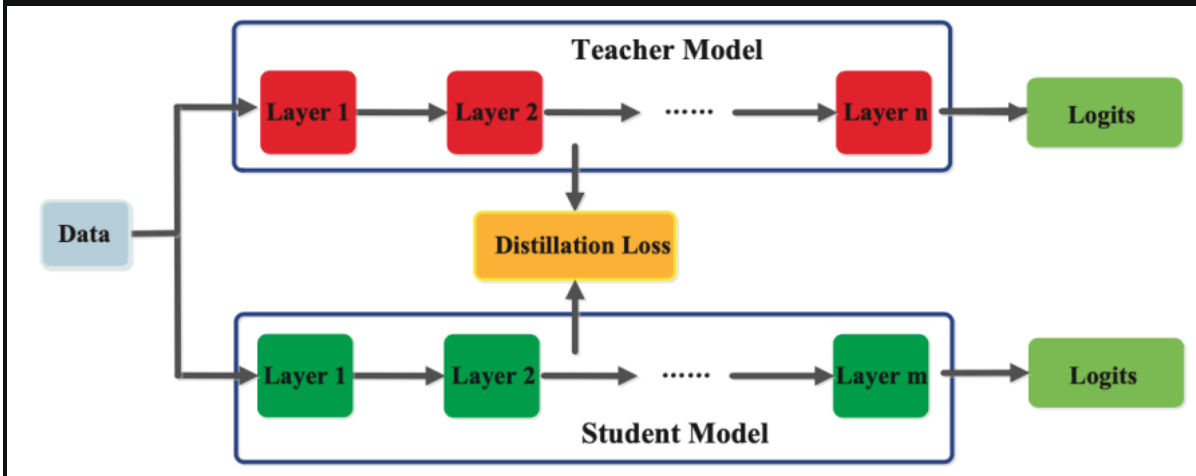

Figure 4. Feature-based knowledge distillation | Source: Arxiv

3. Relation-based knowledge

In addition to knowledge represented in the output layers and the intermediate layers of a neural network, knowledge that captures the relationship between feature maps can also be used to train a student model. This form of knowledge, termed as relation-based knowledge is depicted in Figure 5. This relationship can be modeled as correlation between feature maps, graphs, similarity matrix, feature embeddings, or probabilistic distributions based on feature representations.

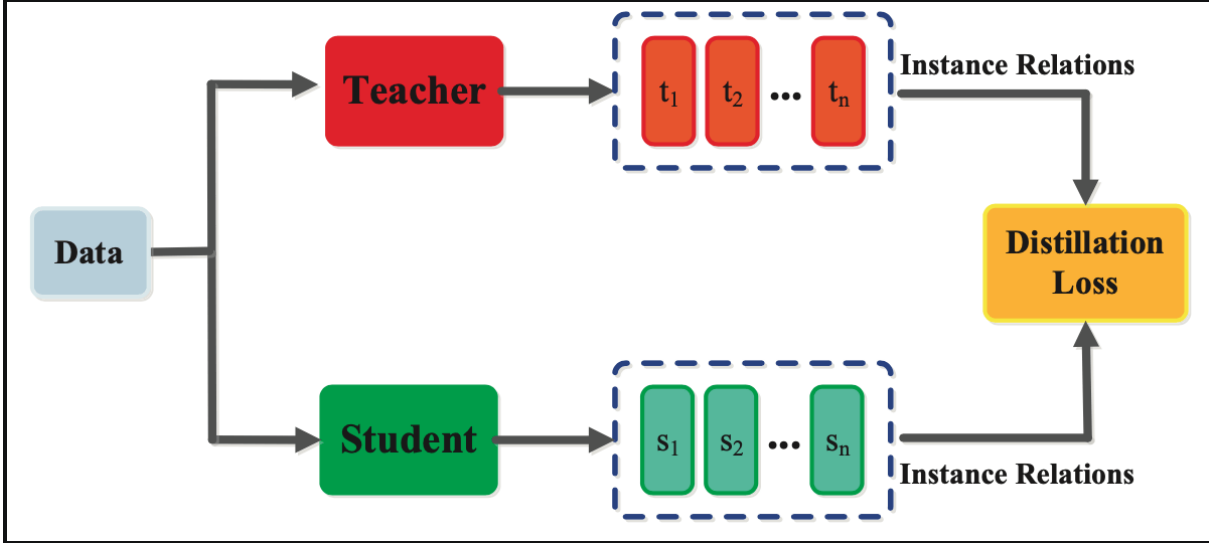

Figure 5. Relation-based knowledge distillation | Source: Arxiv

Training

There are three principal types of methods for training student and teacher models, namely offline, online and self distillation. The categorization of the distillation training methods depends on whether the teacher model is modified at the same time as the student model or not, as shown in Figure 6.

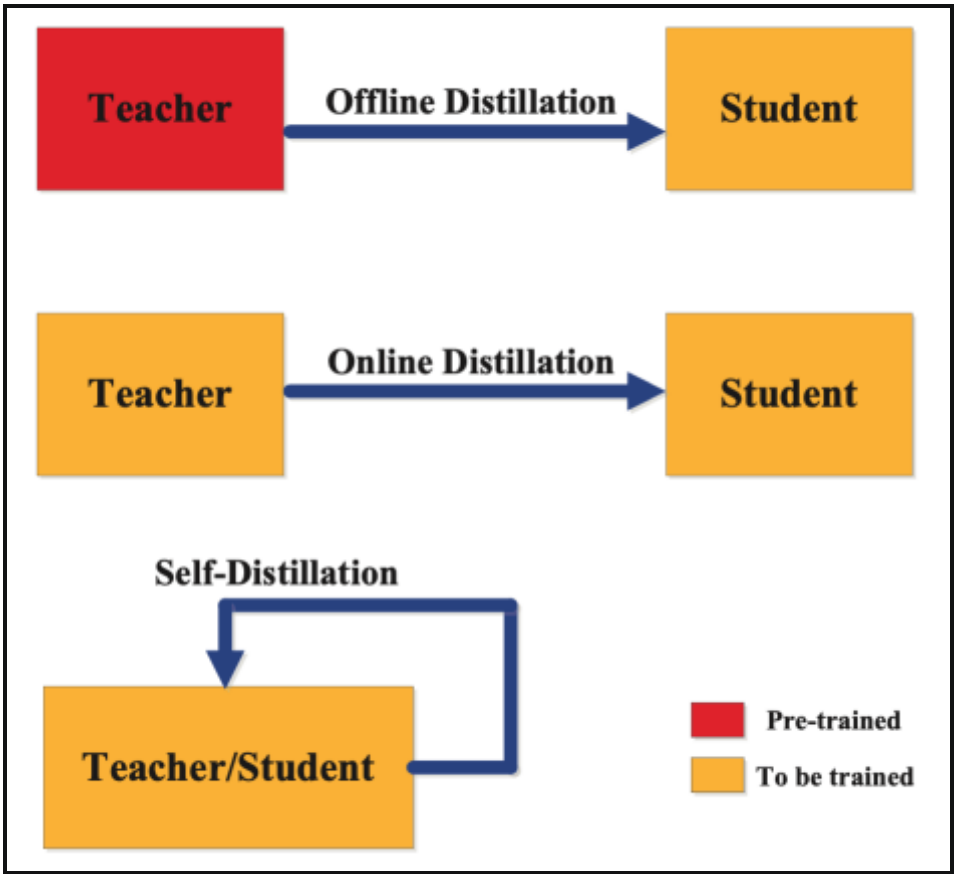

Figure 6. Types of knowledge distillation training schemes | Source: Arxiv

1. Offline distillation

Offline distillation is the most common method, where a pre-trained teacher model is used to guide the student model. In this scheme, the teacher model is first pre-trained on a training dataset, and then knowledge from the teacher model is distilled to train the student model. Given the recent advances in deep learning, a wide variety of pre-trained neural network models are openly available that can serve as the teacher depending on the use case. Offline distillation is an established technique in deep learning and easier to implement.

2. Online distillation

In offline distillation, the pre-trained teacher model is usually a large capacity deep neural network. For several use cases, a pre-trained model may not be available for offline distillation. To address this limitation, online distillation can be used where both the teacher and student models are updated simultaneously in a single end-to-end training process. Online distillation can be operationalized using parallel computing thus making it a highly efficient method.

3. Self-distillation

As shown in Figure 6, in self-distillation, the same model is used for the teacher and the student models. For instance, knowledge from deeper layers of a deep neural

network can be used to train the shallow layers. It can be considered a special case of online distillation, and instantiated in several ways. Knowledge from earlier epochs of the teacher model can be transferred to its later epochs to train the student model.

Architecture

The design of the student-teacher network architecture is critical for efficient knowledge acquisition and distillation. Typically, there is a model capacity gap between the more complex teacher model and the simpler student model. This structural gap can be reduced through optimizing knowledge transfer via efficient student-teacher architectures.

Transferring knowledge from deep neural networks is not straightforward due to their depth as well as breadth. The most common architectures for knowledge transfer include a student model that is:

- a shallower version of the teacher model with fewer layers and fewer neurons per layer,
- a quantized version of the teacher model,
- a smaller network with efficient basic operations,
- a smaller networks with optimized global network architecture,
- the same model as the teacher.

In addition to the above methods, recent advances like neural architecture search can also be employed for designing an optimal student model architecture given a particular teacher model.

Algorithms for knowledge distillation

In this section, I will focus on the algorithms for training student models to acquire knowledge from teacher models.

1. Adversarial distillation

Adversarial learning as conceptualized recently in the context of generative adversarial networks, is used to train a generator model that learns to generate synthetic data samples as close as possible to the true data distribution and a discriminator model that learns to discriminate between the authentic and synthetic data samples. This concept has been applied to knowledge distillation to enable the student and teacher models to learn a better representation of the true data distribution.

To meet the objective of learning the true data distribution, adversarial learning can be used to train a generator model to obtain synthetic training data to use as such or to augment the original training dataset. A second adversarial learning based distillation method focuses on a discriminator model to differentiate the samples from the student and the teacher models based on either logits or feature maps. This method helps the student mimic the teacher well. The third adversarial learning-based distillation technique focuses on online distillation where the student and the teacher models are jointly optimized.

2. Multi-Teacher distillation

In multi-teacher distillation, a student model acquires knowledge from several different teacher models as shown in Figure 7. Using an ensemble of teacher models can provide the student model with distinct kinds of knowledge that can be more beneficial than knowledge acquired from a single teacher model.

The knowledge from multiple teachers can be combined as the average response across all models. The type of knowledge that is typically transferred from teachers is based on logits and feature representations. Multiple teachers can transfer different kinds of knowledge as discussed in section 2.1.

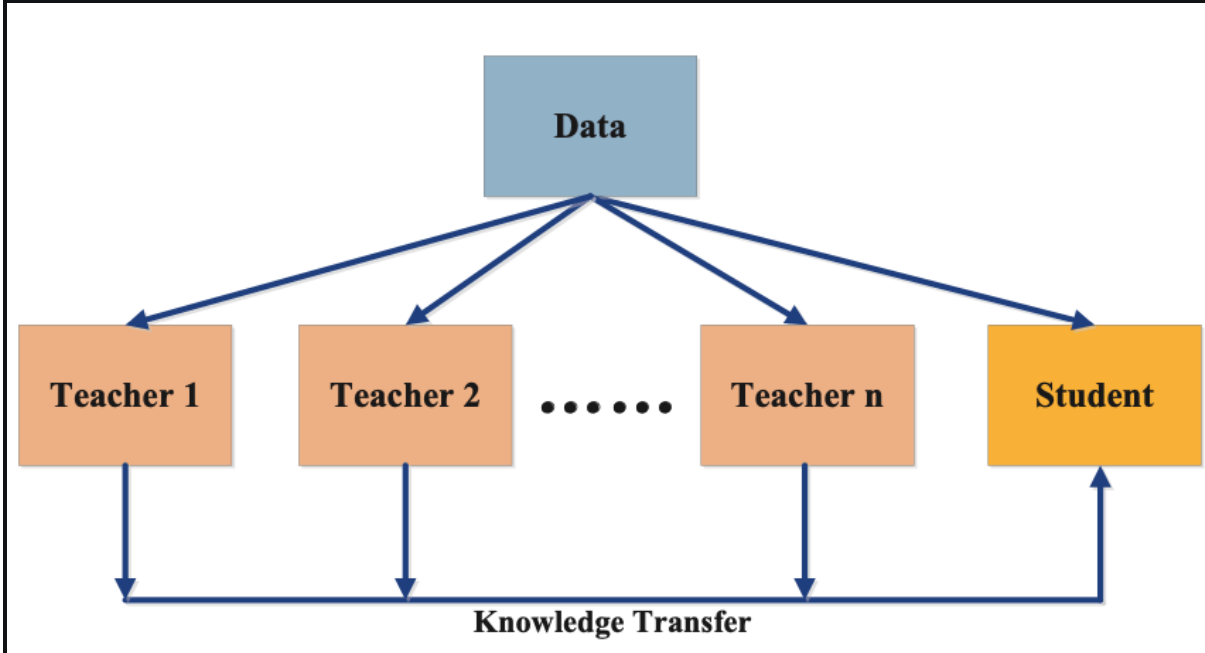

Figure 7. Multi-teacher distillation | Source: Arxiv

3. Cross-modal distillation

Figure 8 shows the cross-modal distillation training scheme. Here, the teacher is trained in one modality and its knowledge is distilled into the student that requires knowledge from a different modality. This situation arises when data or labels are not

available for specific modalities either during training or testing thus necessitating the need to transfer knowledge across modalities.

Cross-modal distillation is used most commonly in the visual domain. For example, the knowledge from a teacher trained on labeled image data can be used for distillation for a student model with an unlabeled input domain like optical flow or text or audio. In this case, features learned from the images from the teacher model are used for supervised training of the student model. Cross-modal distillation is useful for applications like visual question answering, image captioning amongst others.

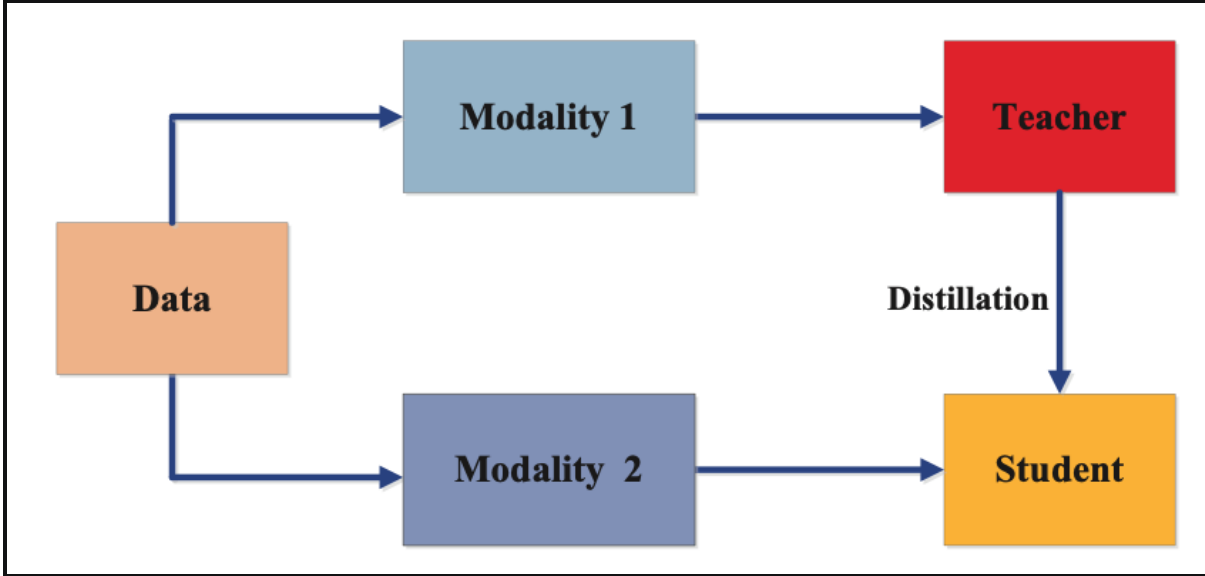

Figure 8. Cross-modal distillation | Source: Arxiv

4. Others

Apart from the distillation algorithms discussed above, there are several other algorithms that have been applied for knowledge distillation.

- **Graph-based distillation** captures intra-data relationships using graphs instead of individual instance knowledge from the teacher to the student. Graphs are used in two ways – as a means of knowledge transfer, and to control transfer of the teacher's knowledge. In graph-based distillation, each vertex of the graph represents a self-supervised teacher which may be based on response-based or feature-based knowledge like logits and feature maps respectively.
- **Attention-based distillation** is based on transferring knowledge from feature embeddings using attention maps.
- **Data-free distillation** is based on synthetic data in the absence of a training dataset due to privacy, security or confidentiality reasons. The synthetic data

is usually generated from feature representations of the pre-trained teacher model. In other applications, GANs are also used to generate synthetic training data.
- **Quantized distillation** is used to transfer knowledge from a high-precision teacher model (e.g. 32-bit floating point) to a low-precision student network (e.g. 8-bit).
- **Lifelong distillation** is based on the learning mechanisms of continual learning, lifelong learning and meta-learning where previously learnt knowledge is accumulated and transferred into future learning.
- **Neural architecture search-based distillation** is used to identify suitable student model architectures that optimize learning from the teacher models.

Applications of knowledge distillation

Knowledge distillation has been successfully applied to several machine learning and deep learning use cases like image recognition, NLP, and speech recognition. In this section, I will highlight existing applications and the future potential of knowledge distillation techniques.

1. Vision

The applications of knowledge distillation in the field of computer vision are plenty. State-of-the-art computer vision models are increasingly based on deep neural networks that can benefit from model compression for deployment. Knowledge distillation has been successfully employed for use cases like:

- image classification,
- face recognition,
- image segmentation,
- action recognition,
- object detection,
- lane detection,
- pedestrian detection,
- facial landmark detection,
- pose estimation,
- video captioning,
- image retrieval,
- shadow detection,
- text-to-image synthesis,

- video classification,
- visual question answering, amongst others [3].

Knowledge distillation can also be used for niche use cases like cross-resolution face recognition where an architecture based on a high-resolution face teacher model and a low-resolution face student model can improve model performance and latency. As knowledge distillation can take advantage of different kinds of knowledge including cross-modal data, multi-domain, multi-task and low-resolution data, a wide variety of distilled student models can be trained for specific visual recognition use cases.

May interest you

Object Detection Algorithms and Libraries
Top Tools to Run a Computer Vision Project

2. NLP

The application of knowledge distillation for NLP applications is especially important given the prevalence of large capacity deep neural networks like language models or translation models. State-of-the-art language models contain billions of parameters, for example, GPT-3 contains 175 billion parameters. This is several orders of magnitude greater than a previous state-of-the-art language model, BERT, which contains 110 million parameters in the base version.

Knowledge distillation is therefore highly popular in NLP to obtain fast, lightweight models that are easier and computationally cheaper to train. Other than language modeling, knowledge distillation is also used for NLP use cases like:

- neural machine translation,
- text generation,
- question answering,
- document retrieval,
- text recognition [3].

Using knowledge distillation, efficient and lightweight NLP models can be obtained that can be deployed with lower memory and computational requirements. Student-teacher training can also be used to address multilingual NLP problems where knowledge from multilingual models can be transferred and shared by each other.

Case study: DistilBERT

DistilBERT is a smaller, faster, cheaper and lighter BERT model [4] developed by Hugging Face. Here, the authors pre-trained a smaller BERT model that can be fine-tuned on a variety of NLP tasks with reasonably strong accuracy. Knowledge distillation was applied during the pre-training phase to obtain a distilled version of BERT model that is smaller by 40% (66 million parameters vs. 110 million parameters) and faster by 60% (410s vs. 668s for inference on the GLUE sentiment analysis task) whilst retaining a model performance that is equivalent to 97% of the original BERT model accuracy. In DistilBERT, the student has the same architecture as BERT and was obtained using a novel triplet loss that combined losses related to language modeling, distillation and cosine-distance loss.

Read also

Unmasking BERT: The Key to Transformer Model Performance

3. Speech

State-of-the-art speech recognition models are also based on deep neural networks. Modern ASR models are trained end-to-end and based on architectures that include convolutional layers, sequence-to-sequence models with attention, and recently transformers as well. For real-time, on-device speech recognition, it becomes paramount to obtain smaller and faster models for effective performance.

There are several use cases of knowledge distillation in speech:

- speech recognition,
- spoken language identification,
- audio classification,
- speaker recognition,
- acoustic event detection,
- speech synthesis,
- speech enhancement,
- noise-robust ASR,
- multilingual ASR,
- accent detection [10].

Case study: Acoustic Modeling by Amazon Alexa

Parthasarathi and Strom (2019) leveraged student-teacher training to generate soft targets for 1 million hours of unlabeled speech data where the training dataset consisted only of 7000 hours of labeled speech. The teacher model produced a probability distribution over all the output classes. The student model also produced a probability distribution over the output classes given the same feature vector and

the objective function optimized the cross-entropy loss between these two distributions. Here, knowledge distillation helped simplify the generation of target labels on a large corpus of speech data.

You might also like

Latent Dirichlet Allocation (LDA) Tutorial: Topic Modeling of Video Call Transcripts (With Zoom)

Conclusions

Modern deep learning applications are based on cumbersome neural networks with large capacity, memory footprint, and slow inference latency. Deploying such models to production is an enormous challenge. Knowledge distillation is an elegant mechanism to train a smaller, lighter, faster, and cheaper student model that is derived from a large, complex teacher model. Following the conceptualization of knowledge distillation by Hinton and colleagues (2015), there has been a massive increase in the adoption of knowledge distillation schemes for obtaining efficient and lightweight models for production use cases. Knowledge distillation is a complex technique based on different types of knowledge, training schemes, architectures and algorithms. Knowledge distillation has already enjoyed tremendous success in diverse domains including computer vision, natural language processing, speech amongst others.

References

[1] Distilling the Knowledge in a Neural Network. Hinton G, Vinyals O, Dean J (2015) NIPS Deep Learning and Representation Learning Workshop. https://arxiv.org/abs/1503.02531

[2] Model Compression. Bucilua C, Caruana R, Niculescu-Mizil A (2006) https://dl.acm.org/doi/10.1145/1150402.1150464

[3] Knowledge distillation: a survey. You J, Yu B, Maybank SJ, Tao D (2021) https://arxiv.org/abs/2006.05525

[4] DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2019) Sanh V, Debut L, Chammond J, Wolf T. https://arxiv.org/abs/1910.01108v4

[5] Lessons from building acoustic models with a million hours of speech (2019) Parthasarathi SHK, Strom N. https://arxiv.org/abs/1904.01624