# 12迁移学习

## 12.1预训练模型

### 12.1.1迁移学习的时代背景？

在当今大数据时代，我们所面临的实际问题中的数据量和数据类型日益增长。这种增长带来了机器学习模型构建和泛化能力方面的新挑战。如何快速构建具有强泛化能力的模型已成为研究的热点之一。而在这个领域，迁移学习作为一种强大的技术，正在被越来越多的人所重视。

在大数据时代中，数据的特点是虽然数据量多，但大多数都没有标注。这意味着，为了训练机器学习模型，我们需要进行大量的标注工作。然而，这种工作需要花费大量的时间和精力，代价是高昂的。因此，利用已有的标签数据和模型进行重用成为了可能。迁移学习就是一种利用已有模型和数据进行模型构建的方法。

传统的机器学习方法通常假定数据服从相同分布。但是，在实际场景中，由于数据来源的差异以及潜在的域偏移问题，很难保证这一假设的正确性。因此，迁移学习方法的出现解决了这个问题。迁移学习可以通过利用源域和目标域之间的相关性，将已有的模型和数据应用到新的领域中。

预训练模型是迁移学习中的一个重要方向。预训练模型在大规模无标注数据上进行训练，从而可以学习到通用的特征表示。这些特征表示可以用于各种任务和领域中的迁移学习。通过预训练模型，我们可以大大减少从头开始构建每一个模型的时间和代价。

### 12.1.2预训练模型的定义和原理

预训练模型是深度学习中的一种重要技术，它通过利用大量未标注数据进行预训练，从而学习出通用的语言知识和特征表示，以适应多个任务的需求。相比于传统的机器学习方法，预训练模型能够更好地处理语言和图像等领域中的复杂问题，取得更好的效果。

预训练模型的基本思想是利用大规模未标注数据进行预训练，然后在有标注数据的任务上进行微调。具体来说，预训练模型通常使用自监督学习方法进行预训练。自监督学习是一种无需人工标注的学习方式，它通过构造一些任务，使得模型在处理数据时需要关注数据的某些特征，从而使得模型能够学习到数据的特征表示。

以计算机视觉为例，预训练模型可以在大规模图像数据集上进行训练，从而学习到图像的通用特征表示，如边缘、纹理等。在下游任务中，可以使用这些预训练模型来初始化卷积神经网络（CNN）的权重，然后再通过微调等技术来训练CNN进行特定任务的学习，如图像分类、目标检测等。这样可以避免从头开始训练模型，节省大量的时间和计算资源，并且通常可以获得更好的性能。

### 12.1.3预训练模型的类别

谷歌在2018年发布的一款NLP预训练模型BERT模型，它采用了独特的神经网络架构Transformer进行语言理解，同时利用双向编码器获取上下文信息，从而有效提高了自然语言处理的表现。BERT模型采用的是掩码语言模型（Masked Language Model，MLM）任务，即在输入序列中随机掩盖一些词汇，让模型预测掩盖的词汇是什么。这个任务可以让模型学习到上下文信息和语言的结构，从而学习到通用的语言知识。BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

预训练模型的优点是能够在大规模的未标注数据上进行预训练，从而学习到通用的语言知识和特征表示，这些特征对于多个任务都有很好的适用性。因此，预训练模型可以在不同领域和任务中进行迁移学习，避免了从头开始训练模型的时间和计算成本。

除了BERT之外，还有许多其他的预训练模型，如GPT、ELMo、RoBERTa、ALBERT等。其中，GPT是一种基于自回归语言模型的预训练模型，它在语言生成和理解等任务上表现出了非常出色的性能；RoBERTa是基于BERT进行改进的预训练模型，采用更大的训练数据和更长的训练时间，取得了更好的效果；RoBERTa: A Robustly Optimized BERT Pretraining Approach;ALBERT是一种轻量级的BERT模型，采用参数共享和嵌入因子分解等技术来减少参数量和计算量，从而在保证性能的情况下减少了模型的大小。ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS这些模型的出现，使得预训练模型在自然语言处理领域得到了广泛应用，同时也促进了深度学习模型的发展和进步。

### 12.1.4预训练模型在迁移学习中的作用

迁移学习是机器学习中的一种重要技术，它可以通过利用已有的模型和数据，来解决新任务中标记数据难以获取的问题。在迁移学习中，预训练模型扮演着重要的角色。

传统的深度学习模型需要大量的标注数据进行训练，这限制了深度学习的应用范围，因为标注数据的获取通常是非常困难和昂贵的。而预训练模型则可以通过在大规模未标注数据上进行预训练，学习到通用的语言知识和特征表示，使得模型更具有泛化能力，可以用较少的标注数据来完成特定任务的学习，进而加速模型的训练过程。

在迁移学习中，预训练模型在迁移学习中通常用于微调和特征提取两个方面：

微调（Fine-tuning）：预训练模型可以作为一个初始化的模型，被用于特定任务的微调。在微调过程中，模型在一个新的任务上用标注数据进行训练，然后根据任务的优化目标来微调预训练模型的参数。这样，模型可以利用预训练模型已经学到的通用特征和知识，同时又能够针对新任务进行优化，从而提高模型的性能。

特征提取（Feature Extraction）：在某些情况下，我们可能只有非常少的标注数据，甚至只有一个小的数据集。这时，我们可以使用预训练模型来提取数据的特征表示，然后使用这些特征表示来训练一个浅层的分类器，以完成特定任务的学习。这种方法通常被称为迁移学习的特征提取方法。这种方法的优点是可以使用预训练模型的通用特征，从而不需要大量的标注数据。

总的来说，预训练模型在迁移学习中的作用是将通用的语言知识和特征表示迁移到新任务上，从而在保留先前学到的知识的同时，又能够针对新任务进行优化，提高模型的性能。

## 12.2微调模型

### 12.2.1什么是微调模型？

在深度学习中，神经网络的低层次的特征通常是较为通用的，因为它们能够检测出基本的边缘、线条、颜色等信息，这些信息在各种数据集中都是相似的。而高层次的特征则更加与具体数据集相关，因为它们的表示更加抽象，能够检测到更加复杂的模式。

在微调模型时，我们通常会冻结预训练模型的底部一些层的参数，即保持这些层的权重不变，不参与更新。因为这些层学习到的是通用特征，可以直接用于新的任务。只有在底部层之上的一些层才需要进行微调，以适应新任务的需求，这样可以在不破坏预训练模型已经学到的特征的情况下，更快地收敛和提高性能。

假设我们想从图像中识别出不同品种的花卉，然后将它们的名称和描述推荐给用户。一种可能的方法是先找出100种常见的花卉，为每种花卉拍摄1,000张不同角度的图像，然后在收集到的图像数据集上训练一个分类模型。如果我们直接使用由ImageNet数据集训练得到的模型，这个花卉数据集的样本数仍然不及ImageNet数据集中样本数的十分之一。这可能会导致适用于ImageNet数据集的复杂模型在这个花卉数据集上过拟合。同时，因为花卉的数据量有限，最终由这些花卉数据集训练得到的模型的精度也可能达不到实用的要求。

为了应对上述问题，一个显而易见的解决办法是收集更多的数据。然而，收集和标注数据会花费大量的时间和资金。例如，为了收集ImageNet数据集，研究人员花费了数百万美元的研究经费。虽然目前的数据采集成本已降低了不少，但其成本仍然不可忽略。

另外一种解决办法是应用迁移学习（transfer learning），将从源数据集学到的知识迁移到目标数据集上。例如，虽然ImageNet数据集的图像大多跟花卉无关，但在该数据集上训练的模型可以抽取较通用的图像特征，从而能够帮助识别边缘、纹理、形状和物体组成等。这些类似的特征对于识别花卉也可能同样有效。

微调模型也是迁移学习的另一个方向。它是在已有的预训练模型的基础上，通过在新的任务上进行训练和调整来获得更好的性能。通过微调模型，我们可以利用预训练模型所学到的通用特征，快速适应新任务并提高性能。

### 12.2.2如何使用微调模型进行迁移学习

在这里，我们要介绍一种经常被用于迁移学习的技术：微调（fine tuning）。微调的步骤包括以下4个阶段：

首先，我们需要在源数据集（如ImageNet数据集）上对神经网络模型进行预训练，得到源模型。

其次，我们需要创建一个全新的神经网络模型，即目标模型。这个模型会复制源模型除了输出层以外的所有模型设计和参数。我们假设这些模型参数已经包含了从源数据集中学习到的知识，并且这些知识同样适用于目标数据集。由于源模型的输出层与源数据集的标签非常相关，因此我们不会在目标模型中使用它。

接着，我们需要在目标模型上添加一个输出层，其大小为目标数据集的类别个数，并随机初始化该层的模型参数。

最后，我们在目标数据集（例如花卉数据集）上训练目标模型。在这个过程中，我们会重新训练输出层，而其他层的参数则是基于源模型的参数微调得到的。

当目标数据集比源数据集小得多时，微调有助于提高模型的泛化能力。

## 12.3领域自适应

### 12.3.1什么是领域自适应？

学会炒菜就更容易学会煲汤、学会数学就更容易学会物理，利用我们已经积累的经验来解决相似的任务是我们的天赋，我们希望人工智能也能具备这种能力。在诸多深度学习领域中存在很多相似的任务，例如计算机视觉领域，无论是区分苹果和橘子、还是区分自行车和摩托车都是图像分类任务，我们希望人工智能能在学习区分苹果和橘子后，利用这段经验来鉴别自行车和摩托车。我们知道卷积神经网络的前几层网络的作用是提取图像特征，最后的全连接层进行分类。于是将训练好后的水果分类网络的前几层直接使用，再使用自行车和摩托车的数据来重新训练最后的全连接层来完成分类。迁移学习就是利用数据、任务、模型间的相似性，将训练好的内容应用到新的任务上。以上两个过程发生在两个数据域之间，其中被迁移的对象被称为源域，被赋予经验的数据域被成为目标域。有时候，目标领域的数据集过小，需要利用源领域的标注数据来辅助完成任务。此外，为了减少训练时间或实现个性化应用，例如疾病诊断、医学图像分析、语言识别、情感分析等，都是迁移学习擅长的领域

### 12.3.2领域自适应的原理

领域自适应是迁移学习的另一个重要方向。在这种情况下，源域和目标域的分布存在差异。通过对源域数据的学习，领域自适应可以帮助模型更好地适应目标域数据的特点。这种方法可以在不需要太多目标域数据的情况下，提高模型的泛化性能。

迁移学习的基本概念

领域自适应（Domain Adaptation）

域（Domain）：数据集的特征分布。

源域（Source domain）：已经有标注数据的域

目标域（Target domain）：没有标注数据的域，需要用源域的数据来迁移学习。

任务（Task）：模型需要完成的具体学习任务，例如图像分类、目标检测等。

条件（Condition）：描述源域和目标域的差异的一些特征，例如不同的拍摄条件、时间和空间分布等。

目标（Objective）：迁移学习的目标，例如在目标领域上获得更好的泛化性能。

限制条件（Constraints）：在迁移学习过程中需要满足的一些限制，例如源领域和目标领域中的数据应该具有一些共同的特征，以便模型能够更好地进行迁移。

在迁移学习中，我们常常需要考虑源域和目标域之间的特征分布情况。如果源域和目标域的特征分布非常相似，那么我们可以很容易地将源域上学到的知识迁移到目标域上，实现迁移学习的目标。相反，如果源域和目标域的特征分布差别很大，我们就需要设计相应的方法来缩小特征分布之间的差异，以达到迁移学习的目的。

### 12.3.3无监督域适应

根据目标域有无标签可以分为一下三类：目标域全部有标签（Supervised DA）、目标域部分有标签（Semi-supervised DA）、目标域全没有标签（unsupervised DA），其中目标域全没有标签（Unsupervised DA）最具有挑战性。

深度学习在不同领域中发挥着重要的作用，其中的原因包括深度学习具有从数据中学习并具有优秀的泛化能力。然而，模型需要大量有标记的数据集进行拟合，训练和测试数据需要满足独立且同分布因此，当在模型适应阶段遇到分布外数据时，深度学习的性能可能会有所下降。为了应对这个问题，提出了无监督领域适应（UDA）

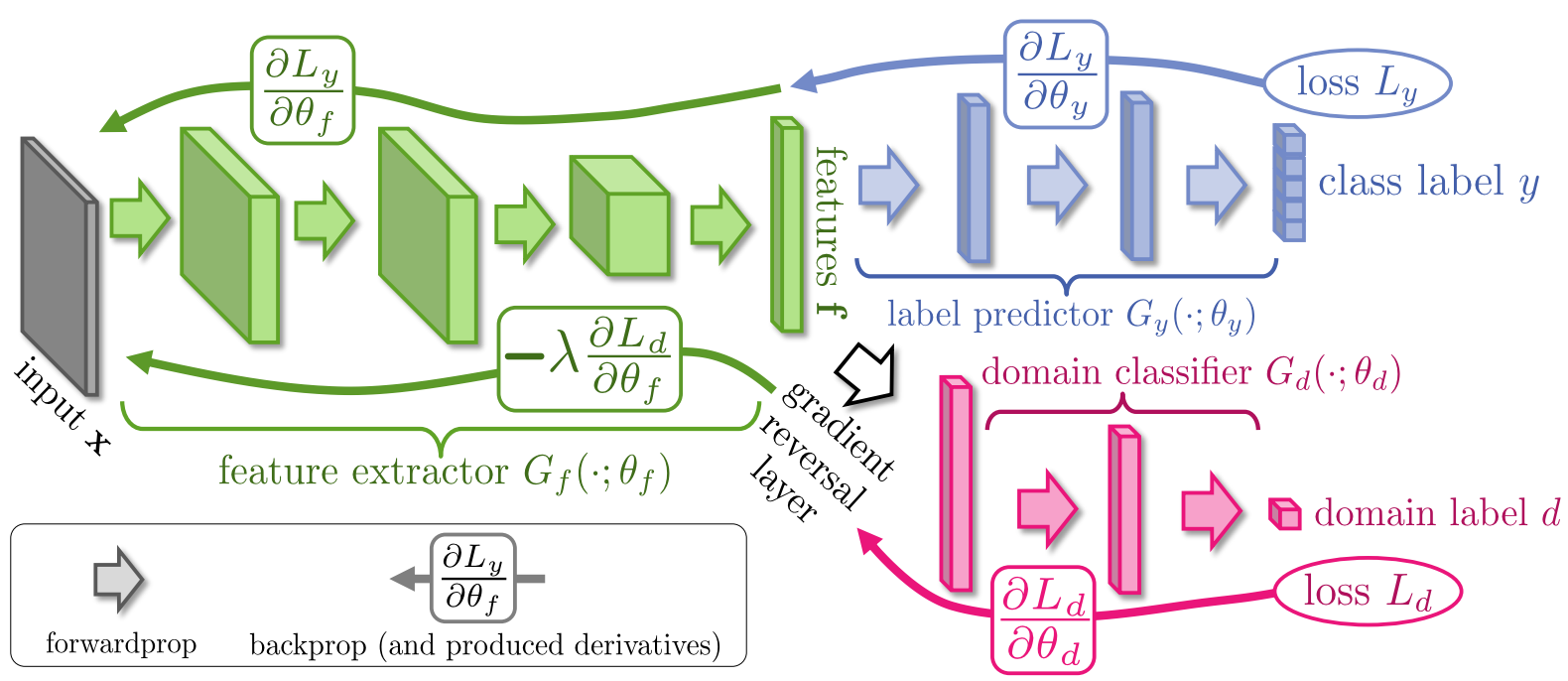
无监督域适应（Unsupervised Domain Adaptation，UDA）是一种域适应的类型，在这种类型中没有标记的目标域数据。在无监督域适应中，目标是通过利用源域数据和未标记的目标域数据中可用的信息学习出在目标域上表现良好的模型。

UDA的基本思想是通过对齐源域和目标域的特征分布来减少域漂移。这可以通过各种技术实现，如对抗训练、领域混淆和自我训练。

一种常见的UDA方法是使用对抗训练，它涉及训练一个域鉴别器，该鉴别器试图区分源域和目标域数据，同时训练一个特征提取器来最小化域分类损失并最大化域混淆损失。另一种方法是领域混淆，它直接在源域和目标域数据之间最大化领域混淆损失。

UDA有许多实际应用，如目标识别、语音识别和自然语言处理，其中往往很难或很昂贵获得标记的目标域数据。通过使用UDA，我们可以利用丰富的标记源域数据和未标记的目标域数据来学习一个能够很好地推广到目标域的模型。

### 12.3.4领域自适应的流程



上图展示了一种架构来实现无监督的领域自适应。该架构由三个部分组成：由特征映射网络Gf(x;sitaf)构成的深度特征提取器（绿色）、由标签分类网络Gy(z;sitay)构成的深度标签预测器（蓝色）和由域判别网络Gd(z;sitad)构成的领域分类器（红色）。深度特征提取器和深度标签预测器一起构成了一个标准的前馈神经网络。在训练过程中，为了实现无监督领域自适应，于是添加了一个领域分类器，并通过一个梯度反转层将梯度乘以某个负常数，从而将特征提取器与领域分类器连接起来。梯度反转的作用是让特征在不同领域上的分布尽可能相似，这样就能够产生领域不变的特征，从而提高模型的泛化能力。在训练过程中，模型将同时最小化标签预测损失和领域分类损失，以实现对不同领域的自适应。[1] Ganin Y , Lempitsky V . Unsupervised Domain Adaptation by Backpropagation:, 10.48550/arXiv.1409.7495[P]. 2014.

其中，源域的数据有标签，而目标域的数据是没有标签的。Gf将源域和目标域的数据都映射到某一个特征空间上，Gy用于预测标签y，而Gd用于预测数据来自于源域还是目标域。因此，带标签的源域数据流入Gy，而不带标签的源域和目标域数据则流入Gd。Gf的作用是将数据映射到特征空间，以便能够分辨出源域数据的标签，而Gd则用于分辨数据来自源域还是目标域。Gy则用于对特征空间的源域数据进行分类，以尽可能分出正确的标签。最终在Gf与Gd相互影响和共同作用下，源域和目标域的数据在特征空间上分布趋于一致，使得Gd无法区分数据来源于源域还是目标域。这样即可使用Gy来对目标域的数据进行分类了。