1. 什么是模型融合？为什么机器学习中模型融合能提高准确率？常见的模型融合方法有哪些？它们的原理是什么？

摘要

深度模型融合/合并是一种新兴技术，它将多个深度学习模型的参数或预测合并为一个模型。它结合了不同模型的能力，以弥补单一模型的偏差和误差，从而实现更好的性能。然而，大规模深度学习模型（如 LLM 和基础模型）的深度模型融合面临着一些挑战，包括高计算成本、高维参数空间、不同异构模型之间的干扰等。尽管模型融合因其在解决复杂现实世界任务方面的潜力而受到广泛关注，但目前仍缺乏对这一技术的完整而详细的调查研究。因此，为了更好地理解模型融合方法并促进其发展，我们对最近的研究进展进行了全面的总结。具体来说，我们将现有的深度模型融合方法分为四个方面：（1）"模式连接"，通过非递增损失路径连接权重空间中的解，从而为模型融合获得更好的初始化；（2）"对齐"，匹配神经网络之间的单元，为融合创造更好的条件；（3）"权重平均"，一种经典的模型融合方法，通过平均多个模型的权重，获得更接近最优解的精确结果。(4) "集合学习 "结合了不同模型的输出，是提高最终模型准确性和鲁棒性的基础技术。此外，我们还分析了深度模型融合所面临的挑战，并提出了未来模型融合可能的研究方向。我们的综述有助于深入理解不同模型融合方法与实际应用方法之间的关联，对深度模型融合领域的研究有所启发。

简介

近年来，深度神经网络（DNN）取得了长足的发展，被广泛应用于计算机视觉（CV）、自然语言处理（NLP）等领域。一般来说，单一的深度学习模型往往存在一定的局限性，无法完全捕捉复杂网络背后的所有底层信息。因此，在深度学习（深度学习）中，经典的集合学习结合了多个模型的输出，以提高模型的最终性能。但它存在测试时存储和运行多个模型的高成本问题，尤其是当模型的复杂度和规模增加时。特别是，例如 GPT-3 有数十亿个参数，PaLM 甚至达到 5400

亿个参数和 7800 亿个令牌。此外，从 DNN 的损失景观来看，梯度优化解通常会收敛到宽平区域边界附近的点，而不是中心点。这意味着经过训练的网络并不完全接近测试误差最小的最优解。为了获得更好的结果，需要对相对最佳点附近的解进行融合。这促使研究人员不仅将融合范围局限于预测（如对数等），还包括在不访问训练数据或维护所有单个模型的情况下融合模型参数。因此，深度模型融合旨在将多个 DNN 融合为一个网络，从而保留其原有能力，甚至优于多任务训练。此外，深度模型融合还能减少单一模型过度拟合特定样本或噪声的倾向，从而提高预测的准确性、多样性和鲁棒性。

由于数据隐私和节省实际资源等问题，深度模型融合引起了越来越多的关注。虽然深度模型融合的发展带来了许多技术突破，但也产生了一系列挑战，如计算负荷高、模型异构、通过组合优化配准速度慢等。有些方法仅限于特定场景，这激发了研究人员对不同情况下模型融合原理的研究。然而，目前还缺乏全面的综述来总结这些方法，从而指出深度模型融合的内部机制。有些研究只关注单一角度的模型融合（如特征融合等）和特定场景的模型融合，或不同方式的信息融合（多模态融合），而非参数融合。为了让开发人员深入了解深度模型融合，我们分析了深度模型融合的原理和方法。此外，我们还回顾了联合学习（FL）和微调等最新进展和代表性应用。我们的调查旨在说明深度模型融合的最新趋势和潜在方向，并为研究人员提高性能和降低成本提供指导。根据内部机制和目的，我们将方法分为四类。对于独立训练的不相邻的模型，"模式连接 "和 "对齐 "可以拉近解的距离，从而获得更好的平均原始条件。对于权重空间存在一定差异的相似模型，"权重平均"倾向于直接对模型进行平均，在损失函数值较低的参数空间区域获得更接近最优点的解。此外，对于现有模型的预测，"集合学习 "会整合不同形式的模型预测，以获得更好的结果。具体来说，可分为以下四类：

模式连通性

基于梯度优化得到的解可以在权重空间中通过一条没有障碍物的路径（连接器）连接起来，这被称为模式连通性。我们可以沿着低损耗路径得到更适合模型融合的其他模型。根据路径的数学形式和连接器所在的空间，我们将本节分为 "线性模式连通性（LMC）"、"非线性模式连通性 "和 "子空间中的模式连通性 "

三个部分。模式连通性可以解决训练过程中的局部优化问题。模式连通性路径的几何关系也可用于加速随机梯度下降（SGD）等优化程序的收敛性、稳定性和准确性。总之，模式连通性为解释和理解模型融合行为提供了一个新的视角。但是，计算复杂性和参数调整等难题亟待解决，尤其是在大型数据集上训练模型时。

对齐

对齐将多个模型的单元进行匹配，并取平均值得到最终模型。对齐后，不同模型之间的特定数学指标（如欧氏距离）可以更加接近，这可以减少模型之间的差异，从而增强深度模型融合的效果。根据是否需要考虑数据分布，配准可分为"激活匹配 "和 "权重匹配"。此外，Re-basin 也是在配准的基础上引入的，它通过排列不变性探索了解可以被传送到单一盆地（即损失相对较小的平坦参数空间区域）的机制。然而，它往往面临计算量大、组合优化速度慢和架构差异等障碍，这使得它不容易扩展到目标不同的其他场景。例如，图匹配带来的内存负担限制了深度模型融合的应用。

权重平均

权重平均是将多个父网络融合为一个网络的最直接、最有效的方法。与模式连接和排列相比，权重平均不需要额外的计算复杂度或训练就能找到优越的起点，在包含一定程度相似性的模型上表现良好。根据聚合空间的不同，权重平均可分为 "权重平均 "和 "子空间平均 "两部分。此外，"模型汤"、"模型算术 "和 "随机加权平均"等典型方法也比现有方法有显著改进。此外，在模型结构或参数数量差异较大的情况下，参数归一化和合并可能会带来一些偏差。尽管如此，权重平均因其简单高效，仍是深度模型融合的主流方法。

集合学习

将多个不同模型的输出结果结合起来，以提高预测性能和鲁棒性，这就是 "集合学习"。在本综述中，我们将重点讨论 深度学习 中的集合学习。在集合学习的基础上，"模型重用 "为每个模型提供了规范，这样在给定新的学习任务时，就能从模型池中识别和合并有用的模型。集合学习有多种框架，界面方便，常用

于物体检测等实际领域。虽然集合学习需要维护多个训练有素的模型，并在测试时运行每个模型，但它仍然是在 深度学习 中被广泛采用的强大技术之一

模型连通性

本节将介绍模式连接的定义、原理和相关方法。在训练神经网络时，基于梯度优化算法（如 SGD 等）训练的解可以合并，而不会产生优越的结果。研究发现，在不增加损失的情况下，解可以通过网络权重空间中的连续路径（连接器）连接起来，这被称为模式连通性。低损耗路径上的模型可以通过模式连通性进行融合，从而发挥多个模型的优势，这对产生更好的聚合模型具有重要意义。

首先，我们解释一下模式连接的原理。在具有代表性的 深度学习 过程中，最小值通常被描述为凸谷底部的一个点，网络参数由最小值的位置决定。传统观点认为，局部极小值和鞍点的数量很大，不同的局部极小值会收敛到参数空间中不同的孤立区域。最近的研究表明，基于梯度的优化器得到的最小值不会被隔离在孤立的山谷中。Gotmare 等人探讨了不同训练过程发现的最小值之间的潜在关系。其他研究表明，神经网络解形成了一个连通的流形（即损失景观中的解在权重空间中通过管道相连）。与模式连通性相比，连接两个独立训练的网络的直接线性路径通常总是留下一个低损耗流形，这就在线性路径上的各点形成了一个高损耗屏障。例如，直接连接两点的线段中点的误差接近 90%（CIFAR-10 上的 VGG-16）。上述工作证明了模式连通性的存在和影响。

其次，一些研究对模式连通性的管道进行了量化。让 $\mathcal{L}(tw_1+(1-t)w_2)$ 为 $t\in(0,1)$ 时的损失（训练或测试误差）。W1 和 W2 之间的线性插值而创建的神经网络的损失（训练或测试误差）。在使用 SGD 的初始化和超参数时，每个历元的随机数据增量可视为噪声。在初始化和超参数固定的情况下使用 SGD 时，每个历元的随机数据增量可视为噪声。为了确定训练后的网络结果是否能稳定地损失屏障（误差屏障）B（w1，w2）被定义为损失屏障和误差屏障线性插值之间的最大差值。的最大差值，如下式所示：

$$B(w_1,w_2)=\sup_t\left[\mathcal{L}(tw_1+(1-t)w_2)\right]-\left[t\mathcal{L}(w_1)+(1-t)\mathcal{L}(w_2)\right].$$

损耗障碍说明了当我们沿着 W1 和 W2 之间的路径优化景观时，误差是不变还是增加。如果两个网络之间有一条屏障近似等于 0 的隧道，这就相当于模式

连通性。也就是说，SGD 得到的局部最小值可以通过最大损耗最小的路径 j 连接起来，如下式所示：

$$\phi\left(w_1, w_2\right)=\underset{\phi \text{ from } \boldsymbol{W}_1 \text{ to } \boldsymbol{W}_2}{\operatorname{argmin}}\left\{\max _{w \in \phi} \mathcal{L}(w)\right\},$$

这意味着通路上的损耗很低，网络对 SGD 噪声很稳定，如图 2 所示。进行模式连通有两个步骤：首先确定隧道的形式（如多边形链、贝塞尔曲线等），如表 1 所示；然后找到连接不同方案的最优低损耗路径，如表 2 所示。根据路径的形式和所处的空间，本节将介绍 ″线性模式连通性″、″非线性模式连通性 ″和 ″子空间中的模式连通性″。

表 1 LMC 和非线性模式连接的标准训练流程汇总

Mode connectivity	The form of path	Ref.	Eq.
Linear path	segment	[54, 58]	$\phi(t)=(1-t) w_1+t w_2$
	polygonal chain	[66, 69]	$\phi(t)=\begin{cases} 2\left(t w+(0.5-t) w_1\right), & 0 \leq t \leq 0.5 \\ 2\left((t-0.5) w_2+(1-t) w\right), & 0.5 \leq t \leq 1 \end{cases}$
Non-linear path	quadratic Bezier curve	[52, 152]	$\phi(t)=(1-t)^2 w_1+2 t(1-t) w+t^2 w_2, \quad 0 \leq t \leq 1$
	Fourier series approximate curves	[234]	$\hat{\phi}(t)=\frac{\beta_0}{2}+\sum_{i=1}^n \beta_i \cos \left(w_i t+\zeta_i\right)$

表 2 找到连接不同方案的最优低损耗路径

Connectors	Methods	Ref.	Introduction
2-dim path	line segment	[71]	produce big error
	GDSS	[61]	approximate the geodesic paths via GDSS
	AutoNEB	[46, 122]	minimize MST to obtain approximation of ϕ^*
	minimize the expectation	[66]	representative approach that connects solutions in a simple way
	RMC	[229]	enhance the robustness of DNNs against different perturbations
N-dim space	MPO	[36, 206]	obtain substantial memory savings
	N-dimensional connectors	[56]	connect low-dimensional wedges
	train parametric subspace	[238]	learn the parameters of lines, curves and simplexes
	SPRO, ESPRO	[11]	find simplexes and simplicial complexes to seek connectors
	geodesic optimization	[215]	speculate the geodesics in the curved distribution space

对齐

表 3 代表性对齐方法的对比

	Alignment	Methods	Ref.
Activation matching	metrics	coefficient of correlation	[140, 218]
		mutual information	[140]
		$\ell 2$ distance	[3, 204, 218]
	pre & post activation	pre-activation	[204, 218]
		post-activation	[140, 218]
Weight matching	metrics	Wassertain distance	[4, 204, 232]
		Euclidean distance	[3, 178]
	graph matching	bipartite matching	[127, 140]
		graph matching	[142]
	other alignment	Bayesian	[227, 254]
		Sinkhorn Re-basin	[178]
		SA	[50]

由于来自不同网络的通道和组件的随机性，网络中的活跃组件会相互干扰。因此，非对齐加权平均可能会忽略不同模型中单元之间的对应关系，从而破坏有用的信息。例如，不同模型中的两个神经元之间可能存在完全不同但功能相似的关系。对齐可以匹配不同模型的单元，从而为深度模型融合提供更好的初始条件。其目的是使多个模型之间的差异更小，从而增强深度模型的融合效果。同时，配准在本质上也可以看作是一个组合优化问题。在本节中，我们将介绍一种具有代表性的机制 "Re-basin"，它可以为各个盆地提供解决方案，从而以更好的原始条件合并模型。之后，我们根据对齐目标是否由数据驱动，将对齐分为 "激活匹配 "和 "权重匹配 "两种类型，如表 3 所示。

权重平均

"权重平均法 "将多个网络的权重结合在一起，以获得性能、鲁棒性和泛化更好的最终模型。它也被称为香草平均法、权重求和法，如下式所示：

$$\sum \lambda_i W_i$$

其中，每个模型都有一个加权参数 λ i，用于控制其对融合模型的贡献程度。然而，与对齐或模式连接不同，WA 的前提条件相对严格。例如，原始模型必须共享部分训练轨迹或位于同一盆地等。这意味着当权重足够相似但有一定差异时，

最终模型可以从所有模型中获益。在平坦的盆地中，解决方案往往表现出良好的性能。反之，狭窄区域的点容易受到能量障碍的影响，导致损失增加。前几节的重点是通过模式连接或排列，将不同区域的解决方案传送到同一盆地。本节将重点讨论同一盆地中解的凸组合的融合，这使得合并后的解更接近盆地的中点（最优点），比端点具有更好的泛化性能，如 SWA、模型汤等。讨论的模型包括以下几种情况：

- 具有一定差异的多个相似模型。
- 对基础模型进行适当微调后的多个模型（如模型汤、模型运算等）。
- 具有相同架构并共享部分训练轨迹的网络的多个检查点（如 SWA、尾平均等）。

集合学习

集合学习或多分类器系统是一种整合多个单一模型以生成最终预测结果的技术，包括投票、平均等。它能提高整体性能，减少模型的方差，解决过拟合、不稳定和数据量有限等问题。在本节中，我们将展示深度学习中的 "集合学习 " 和相关技术 "模型重用"。

集合学习将网络的输出结果结合在一起，比任何一个模型单独学习的结果都要好。一般的平均化模型权重，即 $f\left(x, \frac{1}{n} \sum_{i=1}^{n} W_i\right)$，最终只有一个模型。与此相反，集合学习对推理 $\frac{1}{n} \sum_{i=1}^{n} f\left(x, W_i\right)$ 后的输出值进行平均，从而得到多个模型。集合学习的研究由来已久。有很多典型算法，如 Adaboost、Bagging、Stacking 等。为了让网络表现出更好的泛化能力，之前的一些工作将集合学习（如随机森林等）应用于 DNN，用来调整输出，并充分发挥其在特征选择、噪声过滤方面的优势。Kontschieder 等人提出了深度神经决策森林，在 CNN 的优化算法中使用随机决策函数，以降低参数的复杂度。Zhou 等人介绍了一种决策树集合方法，展示了在没有反向传播的情况下建立模型的可能性，它所需的超参数比典型的深度神经网络要少。

此外，Dropout 通常需要集合所有子网的输出，以减少预测误差。然而，如果多个模型过于相似，不同网络的预测结果就会过于接近，从而无法实现集合学习。为了找到足够多的不同模型，快照集合使用长学习率，将每个学习率周期结束时保存的多个神经网络的预测结果结合起来，产生一个最终结果。作为对快照

的改进，FGE 使用线性片断循环学习率和较小的步长，沿着低损耗路径寻找模型，这启发了 LMC 的相关工作。同样，Laine 等人倾向于将之前多个训练历时的预测集合在一起。Arpit 等人集合了一组包括独立模型和相应的移动平均模型，如下式所示，称为平均集合（EoA）：

$$\hat{y} = \arg\max_{n} \text{Softmax}\left(\sum f\left(x; \hat{W}_i\right)\right)_n$$