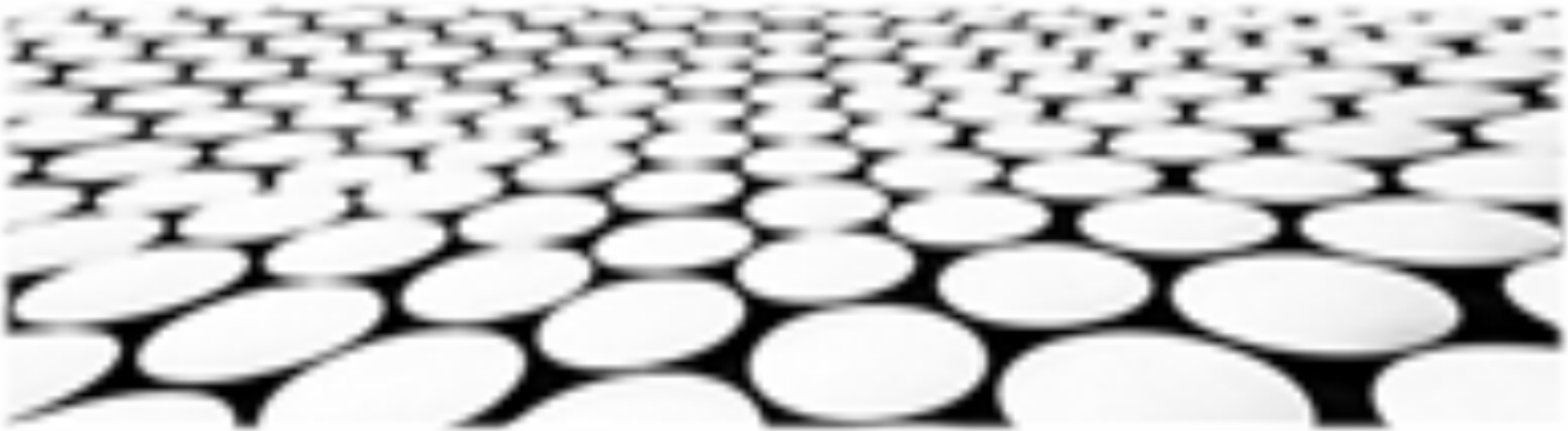
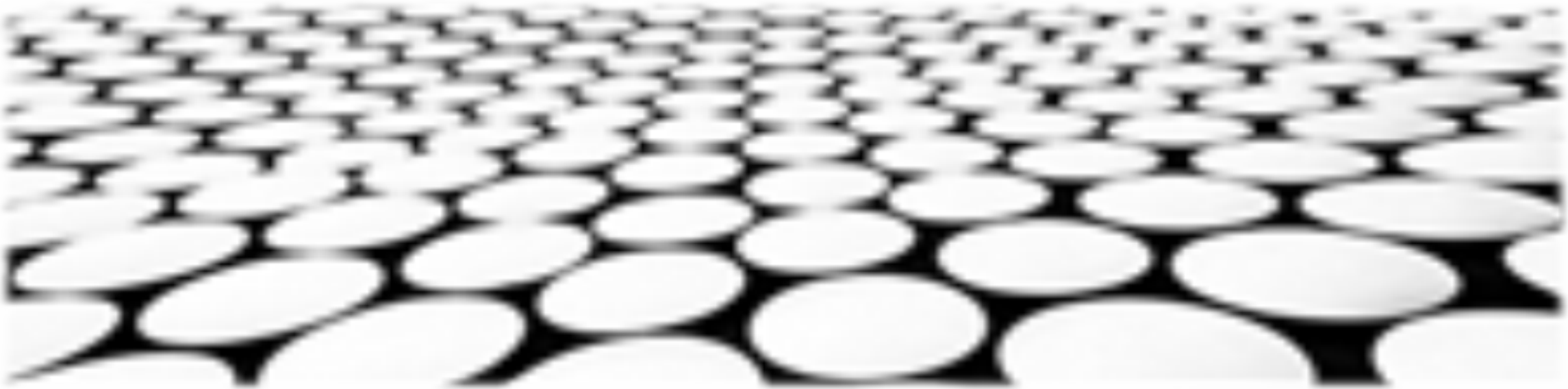
模型压缩与增量更新



目录页

Contents Page

模型压缩概述



模型压缩概述

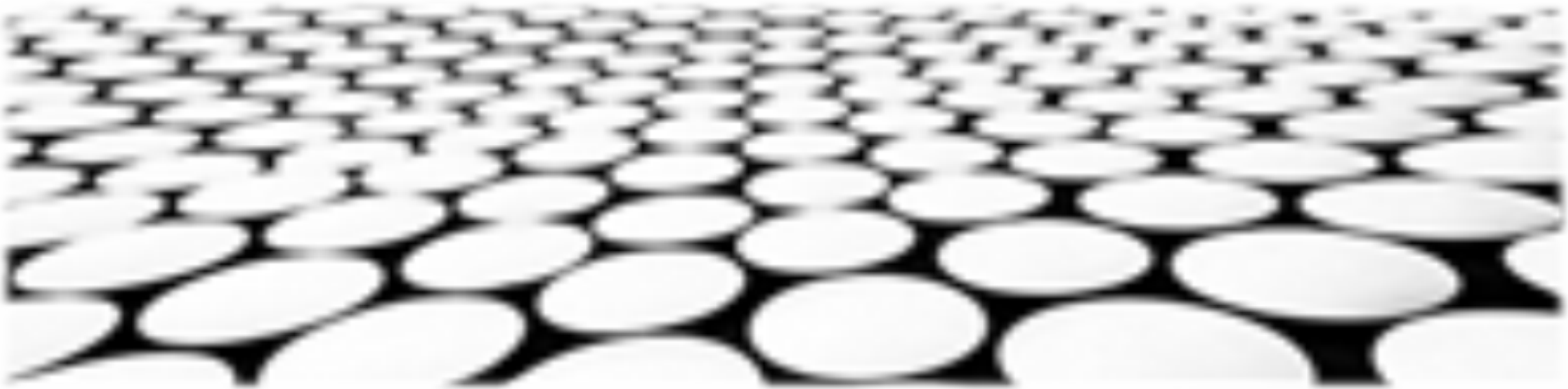
主题名称：模型压缩

1. 模型压缩技术旨在通过减少模型参数和计算复杂度，在保持或提高精度方面进行权衡。
2. 压缩技术包括剪枝、蒸馏、正则化和架构搜索，每种技术都利用了不同的方式来消除冗余和提高效率。
3. 模型压缩对于在资源受限的设备（如移动设备和嵌入式系统）上部署深度学习模型至关重要。

主题名称：模型增量更新

1. 模型增量更新是一种持续学习技术，它涉及逐步更新模型，而不是从头开始重新训练。
2. 增量更新可以节省计算成本、提高适应新的数据并避免漂移。

模型增量更新原则



模型增量更新原则

#模型增量更新原则主题名称：逐步更新

1. 对模型进行分步式更新，一次仅更新模型的一部分。
2. 允许在训练过程中逐步引入新数据或更改，从而避免重新训练整个模型。
3. 减少计算成本和培训时间，特别是在 处理大型数据集或频繁更改时。

主题名称：参数共享

1. 训练多个模型（父模型和子模型）并共享公共参数。
2. 子模型仅更新其特定参数，而父模型的参数保持不变。
3. 减少内存消耗并简化模型部署，同时保持模型性能。

模型增量更新原则

主题名称：知识蒸馏

1. 将训练好的较大的“教师”模型的知识转移到较小的“学生”模型中。
2. 学生模型通过模仿教师模型的输出来学习教师模型的特征和表示。
3. 学习效率高，即使对于新的或有噪声的数据。

主题名称：连续学习

1. 使模型能够在不忘记先前学习知识的情况下适应新数据。
2. 模型通过算法或机制不断调整其参数，以适应变化的环境或任务。
3. 适用于需要持续训练和适应的实时系统或应用程序。

模型增量更新原则

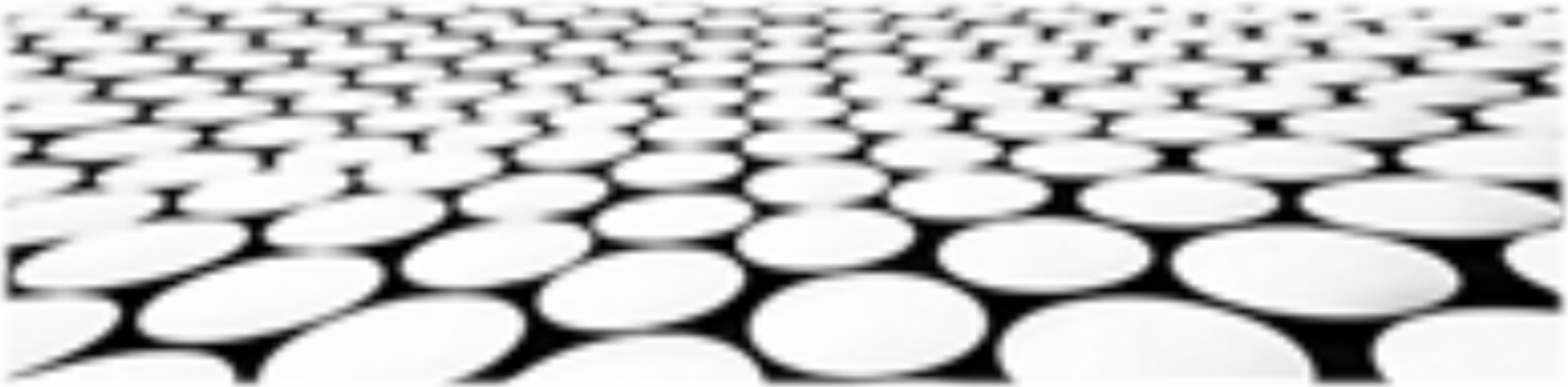
主题名称：弹性更新

1. 允许模型在更新期间具有鲁棒性和容错性。
2. 即使在遇到数据分布变化或部署错误的情况下也能保持模型性能。
3. 通过引入冗余机制或使用元学习技术实现。

主题名称：自动更新

1. 消除手动模型更新的需要，实现模型的自主更新。
2. 基于性能指标或业务规则 自动触发模型更新过程。

剪枝和量化技术



剪枝和量化技术

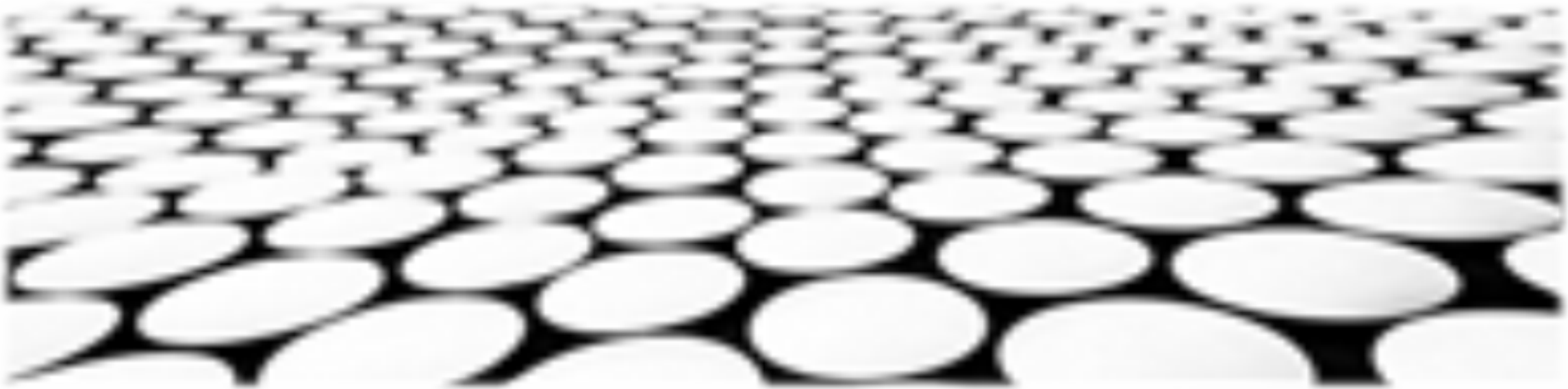
剪枝技术

1. 剪枝技术通过移除不重要的权重和神经元，可以大幅度压缩模型的大小。
2. 剪枝方法通常采用渐进式的方式，通过迭代训练模型并去除权重较小的连接来逐步减小模型规模。
3. 剪枝技术对于降低存储空间、提高推理速度和减少能耗非常有效。

量化技术

1. 量化技术将模型的权重和激活值表示为低精度格式，例如 int8 或 float16，以减少模型的大小。
2. 量化方法需要考虑精度与性能之间的权衡，以确保压缩后的模型不会显著影响模型精度。
3. 量化技术与剪枝技术相结合，可以实现更加有效的模型压缩效果，在保持模型性能的同时最大程度地减小模型大小。

模型膨胀分析



以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：
https://d.book118.com/338036132070006110