合成自适应引导嵌入（SAGE）：一种新型知识蒸馏方法

Suleyman O. Polat Poli A. Nemkova Mark V. Albert
University of North Texas
Denton, USA
{suleymanolcay.polat, poli.nemkova, mark.albert}@unt.edu

2025 年 8 月 22 日

摘要

模型蒸馏能够将大型模型的知识转移到紧凑的学生模型中，从而便于在资源受限的环境中部署。然而，传统的蒸馏方法常常受到计算开销和有限泛化能力的影响。我们提出了一种新颖的自适应蒸馏框架，在学生模型损失较高的区域动态增强训练数据。通过基于 UMAP 的降维和最近邻采样，我们的方法能够识别嵌入空间中的表现不佳区域，并生成有针对性的人工样本以引导学生学习。为了进一步提高效率，我们引入了一个轻量级的师生接口，绕过了教师模型的输入层，实现了对向量化表示的直接蒸馏。在标准 NLP 基准测试上的实验表明，我们的 6600 万参数的学生模型始终能够匹配或超越已建立的基线，在 QNLI 上达到 91.2%，在 SST-2 上达到 92.3%，同时使用更少的训练周期。这些结果突显了损失感知数据增强和向量化蒸馏对于高效有效的模型压缩的前景。

1 介绍

近年来，深度学习模型在从计算机视觉 [10, 13] 到自然语言处理 [3, 4] 的广泛任务中实现了最先进的性能。然而，这些模型通常计算成本高昂且内存密集，使得它们在资源受限的设备上部署具有挑战性。模型蒸馏，引入于 [11]，作为一种有前景的技术出现了，通过将复杂、高容量的“教师”模型的知识转移到更简单的“学生”模型中，同时保留性能。

模型蒸馏已被广泛应用于各种应用中，包括模型压缩 [1]、可解释性 [9] 和隐私保护学习 [6]。若干研究表明，蒸馏在减小模型大小的同时保持准确性方面非常有效，使其成为在边缘计算环境和移动应用程序 [5] 中部署深度学习解决方案的关键技术。尽管取得了这些进展，现有蒸馏方法的效率和泛化能力仍然是活跃的研究领域。

近期的努力旨在通过对抗学习 [21]、无数据蒸馏 [16] 和在线蒸馏框架 [7] 来改进蒸馏过程。然而，这些方法通常会引入显著的训练复杂性，需要额外的资源，或无法使监督信号适应学生模型不断演变的弱点。此外，在医疗和自主系统等安全关键领域中，确保强大的泛化能力仍然是一个挑战。

在这篇论文中，我们介绍了一个新的蒸馏框架，通过自适应地生成学生模型嵌入空间高损失区域的合成数据来提高训练效率和模型泛化能力。我们的方法偏离了传统的静态训练流程，根据学生的性能动态调整训练信号，同时完全在向量空间操作以避免原始文本生成或词级别建模带来的开销。这导致了一种简单而有效的方法，可以使用最少的计算和数据冗余来蒸馏高性能模型。

我们的关键贡献是：

- **一种新型自适应蒸馏管道** 通过识别和增强学生模型表现不佳的区域来动态生成合成训练数据。使用基于 UMAP 的降维，随后进行近似逆变换，我们的方法能够在不依赖文本级生成的情况下实现结构化和高效的增强。
- **向量空间训练策略** 跳过了第一个模型层，直接在中间教师表示上进行操作。这种修改显著减少了标记化和早期变换器层的计算开销，加速了训练同时保持了知识保真度。
- **一个迭代的课程式框架** 根据学生的误差分布变化每轮更新训练分布。此集中反馈循环提高了收敛性，促进了泛化，并减少了训练冗余。
- **GLUE 上的实证结果** 表明我们的蒸馏模型在需要细粒度语义理解的任务（例如，RTE，CoLA）中取得了与强基线模型如 DistilBERT、TinyBERT 和 MiniLM 相当或更优的性能，并且有显著的提升。

2 文献回顾

模型蒸馏是一种将大型复杂 “教师” 模型的知识转移到小型高效 “学生” 模型的技术。此过程旨在保持原始模型的预测性能，同时提高可解释性、效率和部署能力。模型蒸馏广泛应用于各种领域，包括深度学习、可解释人工智能和联邦学习。

2.1 基本概念与方法

模型蒸馏的通用方法。研究人员在 [23] 中提出了一个使用中心极限定理确保学生模型统计可靠性的稳定模型蒸馏通用框架，该框架应用于决策树和符号回归。

理论基础。最近的一项研究引入了一个基于 PAC 学习的理论框架用于蒸馏，该框架定义了复杂模型可以被蒸馏的程度，并刻画了这一过程的计算复杂性 [2]。

可解释人工智能中的应用。蒸馏已被用于提高可解释性，特别是在医疗保健等高风险领域。该研究 [20] 展示了学生模型如何生成对医学代码预测的忠实且可信的解释。

2.2 模型蒸馏技术的发展

快速且准确的蒸馏方法。[8] 的作者介绍了 FAST-DAD 技术，该技术可以高效地将 AutoML 集合蒸馏成简单的模型（如提升树和随机森林），同时保持高准确性，并实现超过 10 倍的速度提升。

另一个研究小组提出了一个自知识蒸馏方法，其中模型通过从自己的预测中学习来提高其泛化能力，而无需外部教师。*自我知识蒸馏*。[22]

数据集蒸馏。研究在 [18] 中探讨了数据集蒸馏，其中知识从大型数据集中转移到小型合成数据集中，以实现与原始数据类似的表现。

2.3 模型蒸馏的应用

联邦学习。在联邦学习中，蒸馏技术能够实现跨去中心化设备的模型稳健聚合。作者 [14] 提出了集成蒸馏技术以克服联邦设置中的模型异质性。

自然语言处理（NLP）。知识蒸馏已被应用于自然语言处理任务，如情感分析和机器翻译，以提高模型效率同时保持预测准确性 [17]。

2.4 挑战

尽管有其优势，模型蒸馏面临几个挑战，包括：

隐私问题：近期研究表明，仅靠蒸馏并不能完全抵御成员推理攻击。[12]

数据效率：数据集大小与模型性能之间的权衡仍然是一个正在进行的研究领域，在数据高效蒸馏技术方面有着令人鼓舞的发展方向。

模型蒸馏已成为一种强大的工具，用于使机器学习模型更加高效和易于解释。最近的研究重点是提高蒸馏效率，扩大其在各个领域的应用，并解决隐私问题。随着该领域的发展，预计将开发出更多通用的和保护隐私的蒸馏技术。

3 方法

本研究采用了一种受 MiniLM[19] 启发的教师-学生训练框架，引入了关键修改以提高效率并实现自适应合成数据生成。主要目标是将大型教师模型的知识蒸馏到更紧凑的学生模型中，同时在嵌入空间中学生表现损失最大的区域动态生成额外的训练实例。这种有针对性的数据增强策略确保学生模型获得更加集中的监督，从而即使参数较少也能提升其泛化能力。

3.1 模型架构与训练设置

学生模型遵循与 MiniLM 相同的架构，但我们通过移除其第一层引入了一个关键的修改。我们不直接处理原始文本，而是利用教师模型的第一层将输入文本转换为 768 维（768D）的向量表示，这些表示随后在训练过程中作为教师和学生模型的输入使用。这种方法确保了两个模型在同一表示空间中操作，同时减少了与分词和嵌入层相关的计算开销。

3.2 初始训练和误差分析

我们以在大规模自然语言语料库——维基百科和 BookCorpus[24]——上进行单个周期的训练开始，以获得学生模型的初始参数化。这一预热阶段提供了粗略的语言先验知识，同时避免了全规模预训练的计算成本。

初始化之后，我们使用我们的蒸馏目标（例如，对数概率的均方误差或软交叉熵）来计算学生模型和教师模型输出之间的实例级损失。我们将示例按此损失进行排序，以识别最具挑战性的样本——那些学生的预测与教师

的输出差异最大的样本。这些“难以学习”的实例暴露了学生模型缺乏泛化能力的地方，并用于指导我们的合成数据增强策略（见图 1）。

专注于这些高损失区域使我们能够将学习集中在学生的最薄弱环节上。这种针对性的、以错误驱动的方法提高了蒸馏效率，加快了收敛速度，并通过避免对学 生已经掌握的例子进行过度训练来减少冗余。

3.3 降维用于数据增强

为了系统地分析模型性能并生成新的训练实例，我们应用一致流形逼近和投影 (UMAP)[15] 将 768D 向量表示降维到 2D 空间。这种降维服务于两个关键目的：

1. **促进最近邻搜索：**在高维嵌入空间（例如，768 维）中，由于维度灾难的影响，传统的距离度量方法如欧几里得或余弦相似度变得不太可靠。随着维度的增加，数据点倾向于等距分布，使得难以有意义地区分相似和不相似的例子。这严重限制了最近邻搜索和聚类的有效性，特别是在尝试识别具有挑战性的高损失样本的连贯区域以进行有针对性的数据增强时尤为如此。为了解决这个问题，我们使用 UMAP 将嵌入投影到一个低维空间——具体来说是 2D，并且使用 100 个邻居。与 PCA 和 t-SNE 相比，UMAP 能够更好地保留局部和全局结构，扩展效率高，并支持近似逆变换——使其成为识别高损失样本的聚类并将它们映射回进行增强的理想选择。这种投影在减少噪声和不相关方差的同时保

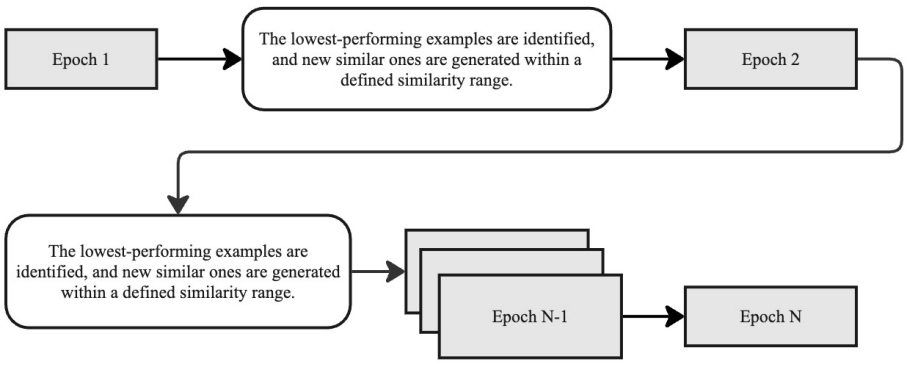

图 1: 训练过程的说明

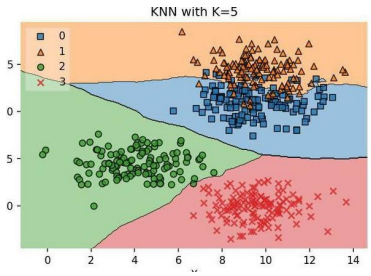

图 2: 使用基于距离的相似性度量生成可比较的挑战性示例：k 近邻

持了局部邻域结构，从而使语义相似、高损失的例子集群更容易区分。在 2D 中操作显著提高了最近邻搜索和聚类的效率和准确性，使我们能够更有效地识别并从特征空间中的表现不佳区域采样。这些检索到的点随后被逆变换回高维空间进行合成数据生成，作为我们的自适应蒸馏框架的一部分。

2. **增强数据多样性：**为了在蒸馏过程中增强合成样本的多样性和泛化能力，我们使用 UMAP 将原始高维嵌入（768 维）投影到 2 维空间。这种降维帮助我们识别学生模型表现出高损失的区域，并允许有针对性的数据扩增。至关重要的是，然后我们执行 UMAP 的大致逆变换，将点从 2D 映射回原始的 768 维空间。虽然这个逆变换并不完全准确，但它对原始数据分布引入了轻微的扰动。这些变化作为一种受控噪声的形式，防止模型过度拟合到原始训练分布，并提高其泛化能力。因此，这一过程作为一种适应性数据扩增形式，针对学生模型的弱点进行定制。

   为了评估 UMAP 投影及其近似逆的保真度，我们进行了定量实验，比较原始和重构的高维向量。我们的结果显示余弦相似性为 0.34，均方误差（MSE）为 0.34，表明虽然逆过程引入了失真，但它仍然保留了足够的结构属性以实现有效的数据增强。这种失真引入的受控扰动有助于训练期间的正则化和改进泛化。

一旦数据被映射到二维，我们使用最近邻算法在高损失区域附近采样新的合成数据点，如图 2 所示。这些合成的二维向量随后通过 UMAP 投影回 768 维，创建出与原始具有挑战性的实例相似但略有变化的新训练样本。

Model	Teacher	#Params	Speedup	Avg
$\text{BERT}_{\text{BASE}}$	-	109M	×1.0	81.5
$\text{RoBERTa}_{\text{BASE}}$	-	125M	×1.0	86.2
$\text{BERT}_{\text{SMALL}}$	-	66M	×2.0	79.1
Truncated $\text{BERT}_{\text{BASE}}$	-	66M	×2.0	76.2
Truncated $\text{RoBERTa}_{\text{BASE}}$	-	81M	×2.0	77.6
DistilBERT	$\text{BERT}_{\text{BASE}}$	66M	×2.0	79.4
TinyBERT	$\text{BERT}_{\text{BASE}}$	66M	×2.0	79.1
MiniLM	$\text{BERT}_{\text{BASE}}$	66M	×2.0	79.4
SAGE	$\text{BERT}_{\text{BASE}}$	66M	×2.0	78.6

表 1: 模型特征和平均性能

Model	SQuAD2	MNLI-m	QNLI	QQP	RTE	SST	MRPC	CoLA
$\text{BERT}_{\text{BASE}}$	76.8	84.5	91.7	91.3	68.6	93.2	87.3	58.9
$\text{RoBERTa}_{\text{BASE}}$	83.7	87.6	92.8	91.9	78.7	94.8	90.2	63.6
$\text{BERT}_{\text{SMALL}}$	73.2	81.8	89.8	90.6	67.9	91.8	88.2	43.3
Truncated $\text{BERT}_{\text{BASE}}$	69.9	81.2	87.9	90.4	65.5	90.8	82.7	41.4
Truncated $\text{RoBERTa}_{\text{BASE}}$	70.9	82.0	89.4	90.5	69.8	92.4	85.6	42.5
DistilBERT	73.3	83.5	90.5	90.8	72.2	91.6	88.5	42.8
TinyBERT	73.1	83.0	90.3	90.5	72.9	91.6	88.3	42.4
Sage	74.2	83.4	91.2	90.8	68.5	92.3	86.9	41.5

表 2: 各任务性能（%）对比模型

3.4 迭代训练和收敛

在每个训练周期中，新生成的合成数据替换之前的数据库，使学生模型能够迭代地改进其表示并更接近地逼近教师模型。这个迭代过程一直持续到学生达到预定义的表现阈值——通常设定为在教师标注的数据集上达到99%的准确性——这通常在十个时期内实现。通过不断调整训练分布以针对学生模型损失最大的区域，该框架确保学习保持专注和高效。这种动态采样

策略不仅加速了收敛，还减少了训练信号中的冗余，从而实现了更有效的知识转移，并且需要较少的训练迭代。

3.5 计算环境

所有训练实验均在使用 A100 GPU 的 Google Colab[1] 上进行。此设置提供了实时降维、合成数据生成和模型微调所需的计算资源，确保了高效的训练周期。

通过将自适应数据增强与教师-学生蒸馏框架相结合，我们的方法在保持计算效率的同时提高了模型的收敛性。基于 UMAP 的难度评估和最近邻合成数据生成的结合使训练过程更加有针对性和可解释。

4 实验

我们选择了 GLUE 基准作为我们的评估框架，因为该框架提供了一系列多样化的自然语言理解任务，这些任务测试了模型泛化能力的各个方面，包括句子相似性、蕴含关系和情感分类。这种多样性使我们能够严格评估精简后的模型在不同领域和任务类型之间的迁移效果。此外，GLUE 是 NLP 社区广泛采用的标准之一，这使得我们可以与现有的学生模型和基线模型（如 DistilBERT 和 TinyBERT）进行有意义的比较。通过在 GLUE 上进行评估，我们确保了我们的方法既具有实际相关性，也具备竞争力。

5 结果与讨论

结果可以在表 2 中找到。我们的模型在与其他基于蒸馏的方法相比时表现出竞争力，在多个自然语言处理基准测试中取得了优异的成绩。值得注意的是，它保留了与 DistilBERT、TinyBERT 和 MiniLM 等其他学生模型相似的 66M 参数规模，表明它在相同的资源约束下运行。

主要观察结果：

- **MNLI-m（83.4%）**：稍低于 MiniLM 和 DistilBERT（83.5%），但与 TinyBERT 相当。这表明我们的模型在蕴含任务上保持了较强的泛化能力。

[1] https://colab.research.google.com/

- **QNLI (91.2%)**：高于 DistilBERT、TinyBERT 和 MiniLM（90.5%），表明在基于问答的自然语言推理方面有所改进。
- **QQP (90.8%)**：在重复问题检测中，表现出与 MiniLM 和 DistilBERT 相当的稳健性能。
- **RTE（68.5%）**：低于 MiniLM 和 TinyBERT（72.2%），但与 BERT 相似 $_{\text{small}}$。这表明该模型可能在处理较小的数据集和更具挑战性的推理任务时会遇到困难。
- **SST-2（92.3%）**：比 DistilBERT（91.6%）和 MiniLM（91.8%）略好，展示了强大的情感分类能力。
- **MRPC (86.9%)**：稍低于 MiniLM（88.7%）和 DistilBERT（88.5%），但在识别释义方面仍具有竞争力。
- **CoLA (41.5%):** 模型中最低的，类似于截断的 $\text{BERT}_{\text{base}}$（41.4%），表明可能在捕捉语言可接受性方面存在弱点。

总体而言，我们的模型在各项任务中表现出良好的均衡性能，在 QNLI 和 SST-2 上尤其出色。然而，其 RTE 和 CoLA 分数表明有改进的潜力，特别是在需要精细语言推理的任务方面。

6 消融研究

6.1 降维的影响维度数量减少

为了更好地理解降维在我们自适应蒸馏框架中的作用，我们进行了一项消融研究，变化 UMAP 投影中使用的维度数量。我们的基线系统使用的是一个 768D→2D→768D 的投影。在这项研究中，我们将比较以下变体：

- **无 UMAP：** 最近邻搜索和合成增强直接在原始的 768 维空间中进行。
- **UMAP-3D:** 嵌入在近似逆之前被投影到 3 维。
- **UMAP-4D:** 投影到 4 维。
- **UMAP-8D:** 投影到 8 维。
- **UMAP-16D:** 投影到 16 维。

降维	平均 GLUE 得分
No UMAP (768D native)	78.1
UMAP-2D (baseline)	**78.6**
UMAP-3D	78.3
UMAP-4D	78.2
UMAP-8D	78.3
UMAP-16D	77.9

表 3: 不同 UMAP 投影维度的平均 GLUE 分数。

- **UMAP-2D（基线）：**我们默认的二维投影设置。

这个受控实验孤立了投影维度的数量如何影响合成增强的质量和下游任务性能。

6.2 结果与分析

所有 GLUE 任务中每种配置的平均性能总结如表 6.2 所示。我们观察到，虽然完全消除降维导致最弱的性能，适度的降维（从 3D 到 8D）提供的性能与 2D 相当或略好。然而，非常高的投影维度或非常低的投影维度倾向于要么欠正则化要么过度扭曲嵌入。

这些发现强化了降维不仅作为计算工具，而且还作为一种归纳偏置，能够增强合成数据的多样性和模型泛化能力的作用。最佳性能是在 **UMAP-2D** 中观察到的。

相比之下，完全移除 UMAP 会导致学生模型训练在过于冗余或无结构的数据上，从而导致性能较弱。这些结果支持我们的假设，即维度感知的合成数据生成在模型效率和泛化能力中起着关键作用。

7 限制和未来工作

虽然我们的方法提高了训练效率和模型性能，但它也存在一些需要进一步考虑的限制。

- 有损降维：将高维嵌入（768 维）压缩到二维空间可能会引入失真，从而可能损害重构的合成数据的保真度和表示质量。

- *计算开销*：包含 UMAP 投影、最近邻检索和高维向量逆变换的迭代管道相比标准蒸馏方法会产生额外的计算成本。
- *教师模型依赖性*：学生模型的有效性仍然受到教师模型的限制和潜在偏差的影响，这可能会限制可实现性能的上限。
- *缺乏文本级可解释性*：由于增强过程完全在向量空间中进行，不将合成示例解码为人类可解释的文本，因此很难评估或验证学生模型正在学习的内容。这可能会妨碍定性分析、错误诊断或与任务语义的一致性，尤其是在关键安全应用中。
- *超越 GLUE 的泛化能力尚未得到证明*：该方法仅在 GLUE 上进行了评估——这是一个研究充分但相对狭窄的基准。其在其他 NLP 基准测试中的表现，包括多跳推理、对话或跨语言任务等，仍未被探索。
- *近似逆可能引入没有结构的噪声*：虽然从 UMAP 逆变换中产生的受控噪声被用作一种数据增强形式，但无法保证重构的向量对应于连贯的语言概念。这可能导致在语义上无效的表示上进行训练，可能会损害边缘情况下的学习。

7.1 未来工作

为了进一步加强和扩展我们的方法，未来的研究方向包括：

- 研究替代或混合的降维技术，以更好地保留语义结构同时最小化信息损失。
- 提高合成数据生成的效率和多样性，可能通过生成建模或对比采样策略实现。
- 将框架扩展到更大的数据集和更复杂的基于变压器的学生架构中，以评估其稳健性和可扩展性。
- 评估在更广泛的 NLP 基准测试中的泛化能力，包括低资源和跨语言任务，以评估方法的迁移性。

解决这些限制对于提高自适应蒸馏方法的可扩展性、灵活性和实际应用性至关重要。

8 结论

我们引入了 **sage**，一种新颖的自适应蒸馏框架，通过结合向量空间训练与目标合成数据增强来提升学生模型的表现。通过 UMAP 识别嵌入空间中的高损失区域，并通过近似逆运算生成扰动训练样本，SAGE 将监督重点放在学生的最薄弱环节上。此外，我们的跳层界面直接在中间教师表示上操作，从而减少计算开销。

在 GLUE 基准测试中的实证结果表明，SAGE 相比 DistilBERT 和 MiniLM 等已建立的蒸馏基线模型，在需要较少训练周期的情况下能够达到具有竞争力或更优的表现。迭代且损失感知的训练范式促进了高效收敛和改进泛化能力，特别是在涉及语义推理和情感分类的任务上。

我们的研究结果强调了嵌入空间引导增强在蒸馏中的有效性，并为可扩展、资源高效的模型压缩开辟了新途径。未来的研究将探索替代的降维技术，进一步多样化合成样本生成，并将框架扩展到多语言和低资源 NLP 设置中。

参考文献

[1] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[2] Enric Boix-Adserà. Towards a theory of model distillation. *ArXiv*, 2024.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, 2018.

[4] Alec Radford et al. Language models are few-shot learners. In *NeurIPS*, 2019.

[5] Jianping Gou et al. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[6] Nicolas Papernot et al. Semi-supervised knowledge transfer for deep learning from private training data. *International Conference on Learning Representations (ICLR)*, 2016.

[7] Yonggang Zhang et al. Deep mutual learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] Rasool Fakoor, Jonas Mueller, Nick Erickson, P. Chaudhari, and Alex Smola. Fast, accurate, and simple models for tabular data via augmented distillation. *ArXiv*, 2020.

[9] Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

[12] Matthew Jagielski, Milad Nasr, Christopher A. Choquette-Choo, Katherine Lee, and Nicholas Carlini. Students parrot their teachers: Membership inference on model distillation. *ArXiv*, 2023.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[14] Tao Lin, Lingjing Kong, S. Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *ArXiv*, 2020.

[15] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[16] Augustin Micaelli and Amaury Labrune. Zero-shot knowledge distillation. In *International Conference on Computer Vision (ICCV)*, 2019.

[17] Monir Yahya Salmony and Arman Rasool Faridi. Bert distillation to enhance the performance of machine learning models for sentiment analysis on movie review data. pages 400–405, 2022.

[18] Tongzhou Wang, Jun-Yan Zhu, A. Torralba, and Alexei A. Efros. Dataset distillation. *ArXiv*, 2018.

[19] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

[20] Zach Wood-Doughty, Isabel Cachola, and Mark Dredze. Model distillation for faithful explanations of medical code predictions. 2022.

[21] Haichao Xu and Kristian Kersting. Training a binary neural network from scratch with adversarial learning. In *International Conference on Machine Learning (ICML)*, 2018.

[22] Kai Xu, Lichun Wang, Huiyong Zhang, and Baocai Yin. Self-knowledge distillation with learning from role-model samples. pages 5185–5189, 2024.

[23] Yunzhe Zhou, Peiru Xu, and G. Hooker. A generic approach for reproducible model distillation. *ArXiv*, 2022.

[24] Yukun Zhu. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.06724*, 2015.