Computer Science and Application 计算机科学与应用, 2026, 16(2), 251-260
Published Online February 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.162056

Hans汉斯

知识蒸馏中损失函数的研究进展综述

赵彤彤

河北地质大学信息工程学院，河北 石家庄

收稿日期：2026年1月2日；录用日期：2026年2月2日；发布日期：2026年2月10日

摘　要

知识蒸馏作为一种高效的模型压缩与知识迁移技术，其性能的核心决定因素之一是损失函数的设计。损失函数定义了学生模型模仿教师模型时所遵循的优化目标与知识迁移的维度。本文系统综述了知识蒸馏领域损失函数的研究进展。首先，介绍了基于输出响应的经典损失函数，如KL散度与均方误差。其次，梳理了基于中间层特征匹配的损失函数，包括注意力转移与Hint Learning等方法。接着，总结了基于关系与结构化知识匹配的前沿损失函数，如相似性保持与相关性一致性损失。最后，对知识蒸馏损失函数的研究趋势进行了展望，指出自适应损失组合、面向特定任务的定制化损失以及理论分析是未来的重要方向。本文旨在为研究者，特别是工程应用者，在选择与设计知识蒸馏损失函数时提供一个清晰的参考。

关键词

知识蒸馏，损失函数，模型压缩，神经网络，机器学习

A Review on Research Advances of Loss Functions in Knowledge Distillation

Tongtong Zhao

School of Information Engineering, Hebei GEO University, Shijiazhuang Hebei

Received: January 2, 2026; accepted: February 2, 2026; published: February 10, 2026

Abstract

Knowledge distillation, as an efficient technique for model compression and knowledge transfer, relies critically on the design of its loss functions, which define the optimization objectives and the dimensions of knowledge transfer for the student model to mimic the teacher. This paper provides a systematic survey of the research progsress on loss functions in knowledge distillation. Firstly, it introduces classical loss functions based on output responses, such as Kullback-Leibler divergence and mean squared error. Secondly, it reviews loss functions based on intermediate feature

文章引用: 赵彤彤. 知识蒸馏中损失函数的研究进展综述[J]. 计算机科学与应用, 2026, 16(2): 251-260.
DOI: 10.12677/csa.2026.162056

matching, including attention transfer and hint learning. Subsequently, it summarizes advanced loss functions based on relational and structured knowledge matching, such as similarity-preserving and correlation congruence losses. Finally, future research trends are discussed, pointing out that adaptive loss combination, task-specific customization, and theoretical analysis are important directions. This paper aims to provide a clear reference for researchers, especially practitioners, in selecting and designing loss functions for knowledge distillation.

Keywords

Knowledge Distillation, Loss Function, Model Compression, Neural Network, Machine Learning

1. 引言

深度神经网络在计算机视觉、自然语言处理等领域取得了突破性成功，但其庞大的参数量与计算成本限制了在资源受限环境(如移动设备、嵌入式系统)中的部署[1] [2]。知识蒸馏作为一种有效的模型压缩与加速技术应运而生，其核心思想是训练一个轻量级的“学生”网络，使其不仅学习原始数据标签，更重要的模仿一个预先训练好的、性能强大但复杂的“教师”网络的内部知识或行为，从而在保持较高性能的同时显著降低模型复杂度[3] [4]。

在知识蒸馏的框架中，损失函数是连接教师网络与学生网络的桥梁，它决定了学生网络需要模仿教师网络中的何种知识，以及如何量化这种模仿的差距。损失函数的设计直接关系到知识迁移的效率和最终学生模型的性能。自Hinton等人提出基于软化输出的知识蒸馏以来，研究者们从不同角度对损失函数进行了广泛而深入的探索，其演进路径清晰地反映了对“知识”定义的理解从浅层到深层的深化过程。

最初的损失函数聚焦于匹配教师与学生网络的最终输出(即“结果”)。随后，研究者认识到中间层特征蕴含了丰富的结构化信息，于是催生了基于特征匹配的损失函数[5]。此后，研究进一步深化，例如通过注意力转移(Attention Transfer)将“知识”提炼为特征图的空间注意力图，强调迁移“哪里重要”的空间结构信息，而非具体的特征值，实现了对特征知识更精炼的抽象[6]。近年来，前沿研究更进一步，深入到挖掘样本间或特征通道间的关系知识，并设计了相应的关系匹配损失[7]。尽管这些损失函数形式多样，但其数学核心大多可归结为交叉熵/KL散度、均方误差(L2)和绝对误差(L1)等基础度量在不同知识维度上的应用。

本文旨在对知识蒸馏中损失函数的研究进展进行系统性梳理。首先，回顾基于输出响应的经典损失函数。其次，总结基于特征匹配的损失函数及其变体。然后，介绍基于关系与结构化知识匹配的前沿损失函数。最后，对当前研究的局限进行讨论，并对未来可能的发展方向进行展望，以期为相关领域的研究者与应用者提供参考。

2. 基于输出响应的损失函数

基于输出响应的匹配是知识蒸馏最经典、最直接的形式。这类方法的核心思想是使学生网络的最终输出尽可能接近教师网络的输出，其知识定义在网络预测的“结果”层面。

2.1. KL 散度与软化标签

Hinton等人提出的原始知识蒸馏框架使用Kullback-Leibler散度作为核心损失函数[4]。其关键创新在

于引入了“温度”参数 $\mathcal{T}$ 来软化教师网络的Softmax输出，从而产生一个携带更多类间相似性信息的概率分布(即“软化标签”)。学生网络则通过最小化自身软化输出与教师软化输出之间的KL散度来学习。

给定教师网络的logits向量 $\boldsymbol{z}_t$ 和学生网络的logits向量 $\boldsymbol{z}_s$，软化概率分布计算如下：

$$p_i^{\mathcal{T}} = \frac{\exp\left(z_i/\mathcal{T}\right)}{\sum_j \exp\left(z_j/\mathcal{T}\right)} \tag{1}$$

其中，i 表示类别索引，$\mathcal{T}>0$ 为温度参数。当 $\mathcal{T}=1$ 时，即为标准Softmax；当 $\mathcal{T}$ 增大时，产生的概率分布更加平滑。知识蒸馏损失定义为：

$$\mathcal{L}_{\mathrm{KD}} = \mathcal{T}^2 \cdot D_{\mathrm{KL}}\left(\boldsymbol{p}_t^{\mathcal{T}} \,\|\, \boldsymbol{p}_s^{\mathcal{T}}\right) = \mathcal{T}^2 \cdot \sum_i p_{t,i}^{\mathcal{T}} \log \frac{p_{t,i}^{\mathcal{T}}}{p_{s,i}^{\mathcal{T}}} \tag{2}$$

其中，$\mathcal{T}^2$ 项用于确保梯度大小在温度变化时保持相对稳定。总的训练损失通常结合蒸馏损失和标准的交叉熵损失(使用真实硬标签)：

$$\mathcal{L}_{\mathrm{total}} = \lambda_{kd}\mathcal{L}_{\mathrm{KD}} + \lambda_{ce}\mathcal{L}_{\mathrm{CE}}\left(\boldsymbol{p}_s, \boldsymbol{y}\right) \tag{3}$$

其中，λ 为平衡两种损失的权重系数，$\boldsymbol{y}$ 为真实标签。

2.2. 均方误差与直接回归

除了匹配概率分布，一种更直接的方式是让学生网络的logits值直接回归教师网络的logits值。这通常通过均方误差损失实现：

$$\mathcal{L}_{\text{MSE-logits}} = \frac{1}{n}\left\|\boldsymbol{z}_s - \boldsymbol{z}_t\right\|_2^2 \tag{4}$$

其中，n 为向量的维度。这种方法避免了Softmax计算和温度调参，在某些场景下简单有效。其背后的假设是，logits值的相对大小本身就包含了知识。

2.3. 交叉熵与教师标签监督

也有一些工作探索使用教师网络的预测(硬标签或软化标签)作为监督信号，直接使用交叉熵损失训练学生网络[3]。虽然形式简单，但在教师网络置信度很高时，这种方法与使用真实标签训练差异不大，知识迁移的效果有限。因此，它常作为其他更复杂损失的一个组成部分。

Table 1. Comparison of main loss functions based on output response
表 1. 基于输出响应的主要损失函数对比

Brown 损失函数	核心思想与公式	主要特点
KL 散度(软化)	$\mathcal{L}_{\mathrm{KD}} = \mathcal{T}^2 \cdot \sum_i p_{t,i}^{\mathcal{T}} \log \frac{p_{t,i}^{\mathcal{T}}}{p_{s,i}^{\mathcal{T}}}$	传递类间相似性信息，是经典方法，需调节温度 $\mathcal{T}$ 。
均方误差(Logits)	$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{n}\left\|\boldsymbol{z}_s - \boldsymbol{z}_t\right\|_2^2$	直接回归，简单高效，无需 Softmax 和温度参数。
交叉熵(硬标签)	$\mathcal{L}_{\mathrm{CE}} = -\sum_i y_i \log\left(p_{s,i}\right)$	使用教师预测作为监督信号，简单但迁移知识有限。
交叉熵(软化标签)	$\mathcal{L}_{\text{CE-soft}} = -\sum_i p_{t,i}^{\mathcal{T}} \log\left(p_{s,i}\right)$	直接以教师软化概率为监督，可视为 KL 散度的简化变体。

表1总结了基于输出响应的主要损失函数。这类方法的优点是概念直观、实现简单。然而，它们仅利

用了网络的最终输出，忽略了深度神经网络中间层所蕴含的丰富特征表示和结构化信息，这在一定程度上限制了知识迁移的潜力和学生网络性能的上限。

3. 基于中间层特征的损失函数

为了迁移更丰富的知识，研究者提出了基于中间层特征匹配的方法。其核心思想是：教师网络中间层的特征图(Feature Map)包含了输入数据的抽象表示和空间/通道结构信息，引导学生网络中间层的特征图与教师网络的对齐，可以传递更深层次的知识。

3.1. Hint Learning 与回归损失

Romero等人提出的“Hint Learning”是这一方向的先驱工作[5]。该方法指定教师网络的某个中间层为“提示层”，学生网络对应的层为“引导层”。为了匹配两者可能不同的特征图尺寸，引入一个可学习的回归器(通常是一个卷积层)将学生特征投影到教师特征空间。损失函数通常采用均方误差：

$$\mathcal{L}_{Hint} = \frac{1}{WHC}\left\| r\left(\boldsymbol{F}_s\right) - \boldsymbol{F}_t \right\|_2^2 \tag{5}$$

其中，$\boldsymbol{F}_s$ 和 $\boldsymbol{F}_t$ 分别代表学生引导层和教师提示层的特征图，W, H, C 分别为特征的宽度、高度和通道数，$r(\cdot)$ 为回归器。这种方法强迫学生网络学习与教师网络相似的中层特征表示。

3.2. 注意力转移

Zagoruyko等人认为，特征图中空间位置的重要性(即“注意力”)是更值得迁移的知识[6]。他们提出了注意力转移方法，首先通过计算每个空间位置所有通道的 p 范数来生成一个二维的注意力图：

$$A = \left(\frac{1}{C}\sum_{c=1}^{C} \left|F_c\right|^p \right)^{\frac{1}{p}} \tag{6}$$

其中，F_c 是特征图的第 c 个通道，通常取 $p=2$。然后，使用均方误差损失让学生网络的注意力图 A_s 匹配教师网络的注意力图 A_t：

$$\mathcal{L}_{AT} = \frac{1}{WH}\left\| Q\left(A_s\right) - Q\left(A_t\right) \right\|_2^2 \tag{7}$$

其中，$Q(\cdot)$ 是一个归一化函数(如除以所有位置的和)，以确保数值稳定。注意力转移成功地将知识迁移的重点从具体的特征值转移到了特征的空间重要性分布上。

3.3. 特征图的多维度匹配

除了在空间维度上迁移注意力，后续工作探索了在通道维度等其他维度进行匹配。例如，通过计算通道注意力(如SENet中的机制)并进行匹配，可以传递不同特征通道重要性的知识[8] [9]。也有方法将特征图展开为一维向量后，计算其Gram矩阵(风格迁移中常用)并进行匹配，以传递特征通道间相关性的风格信息[10]。这些方法的损失函数通常也采用MSE或L1损失。

基于特征匹配的损失函数使学生网络能够学习教师网络的内部特征表示，从而通常比仅匹配输出获得更好的性能。然而，如何选择匹配的层、如何处理层间不匹配以及如何设计更有效的特征知识度量，仍然是开放的问题。

4. 基于关系与结构化知识的损失函数

近年来，知识蒸馏的研究前沿进一步深入到挖掘和迁移样本之间或特征内部的结构化关系知识。这

类方法认为，教师网络所学到的样本间相似性、特征通道间依赖性等高阶关系，是比单个样本的输出或特征更本质、更鲁棒的知识。

4.1. 相似性保持

相似性保持方法的核心是：对于一个批次内的样本，教师网络在其特征空间(可以是中间层特征或输出层)中形成的样本对相似性关系，应该被学生网络保留[7]。首先，计算教师网络和学生网络中所有样本对之间的相似性矩阵 $\boldsymbol{S}^t$ 和 $\boldsymbol{S}^s$。相似性度量可以是余弦相似性、点积或欧氏距离的负值等。然后，使用损失函数(如MSE或KL散度)最小化两个相似性矩阵的差异：

$$\mathcal{L}_{SP} = \frac{1}{N^2}\left\|\boldsymbol{S}^t - \boldsymbol{S}^s\right\|_F^2 \tag{8}$$

其中，N 为批次大小，$\|\cdot\|_F$ 为Frobenius范数。这种损失迫使学生网络学习教师网络对样本间关系的理解，具有很强的可解释性，并且对输入的变化更加鲁棒。

4.2. 相关性一致性

相关性一致性方法聚焦于特征通道间的关系[11]。它计算教师网络和学生网络特征图中所有通道对之间的相关性矩阵(如皮尔逊相关系数)。该相关性矩阵反映了不同特征通道在响应模式上的统计依赖性。损失函数定义为两个相关性矩阵之间的MSE：

$$\mathcal{L}_{CC} = \frac{1}{C^2}\left\|\boldsymbol{R}^t - \boldsymbol{R}^s\right\|_F^2 \tag{9}$$

其中，$\boldsymbol{R}^t$ 和 $\boldsymbol{R}^s$ 分别为教师和学生的通道相关性矩阵，C 为通道数。这种方法迁移了特征通道间的结构化共生模式，被证明在多种视觉任务上非常有效。

4.3. 对抗式蒸馏中的关系匹配

在基于对抗生成思想的知识蒸馏中，判别器的目标通常是区分特征来自教师还是学生，其损失函数本质上是二元交叉熵[12]。然而，在此框架下，也可以通过让学生网络的特征分布与教师网络的特征分布在更细粒度的关系上(如通过对比学习构建的正负样本对关系)保持一致，来引入关系知识[13]。

基于关系匹配的损失函数代表了当前知识蒸馏研究的一个前沿方向。它们不再关注点对点的知识迁移，而是关注知识体系内部的结构，这使得迁移的知识更加抽象和鲁棒，往往能带来显著的性能提升，尤其是在学生网络与教师网络结构差异较大时。

5. 前沿进展：面向特定架构与目标的损失函数

随着神经网络架构的快速演进(如Vision Transformer，ViT)与模型规模的急剧扩大(如大语言模型，LLM)，知识蒸馏的研究焦点进一步转向针对特定架构设计更高效的损失函数，并深入探索损失函数本身的优化。本节将介绍其中几个代表性方向。

5.1. 解耦知识蒸馏与 Logit 标准化

经典知识蒸馏(KD)损失函数将教师的输出视为一个整体进行模仿。解耦知识蒸馏指出，将教师输出logits $\boldsymbol{z}_t$ 解耦为目标类部分 z_t^c 与非目标类部分 $\boldsymbol{z}_t^{-c}$，并分别进行不同强度的知识迁移，能获得更好的效果。其核心思想是目标类logit提供了最强的类别信号，而非目标类logits (尤其是“难负类”)蕴含了更丰富的类间关系。典型的解耦损失函数设计如下：

$$\mathcal{L}_{\mathrm{DKD}}=\alpha\cdot\underbrace{-\log\frac{\exp\left(z_s^c/\tau\right)}{\sum_i\exp\left(z_s^i/\tau\right)}}_{\text{目标类知识}}+\beta\cdot\underbrace{\sum_{i\neq c}\phi\left(p_t^i,p_s^i\right)}_{\text{非目标类知识}} \tag{10}$$

其中 $\phi(\cdot)$ 为衡量非目标类分布差异的函数。这种方法使学生能更聚焦于学习教师的关键判别知识，已在图像分类、目标检测等任务上展现出优越性[14]。

另一方面，Logit标准化(Logit Standardization)技术旨在解决传统知识蒸馏中共享温度参数所带来的局限性。经典KD假设教师与学生共享同一温度参数，这隐含地要求二者的logits在数值范围与方差上严格匹配。然而，由于模型容量差异，这种强约束限制了学生模型的性能，因为学生只需学习教师logits的内在关系而非精确的数值大小。Logit标准化提出对logits进行Z-score预处理，并将温度设置为logits的加权标准差。具体而言，对教师和学生的logits分别进行标准化：

$$\mathcal{Z}(\boldsymbol{z};\tau)=\frac{\boldsymbol{z}-\mu(\boldsymbol{z})}{\tau},\ \tau=\sqrt{\frac{\sum_{i=1}^{K}\left(z_i-\mu(\boldsymbol{z})\right)^2}{K-1}} \tag{11}$$

其中 $\mu(\boldsymbol{z})=\frac{1}{K}\sum_{i=1}^{K}z_i$ 为logits均值，K 为类别数。经过Z-score变换后，教师与学生的logits具有零均值和单位方差，使学生能够聚焦于学习教师的类间相对关系，而非强制匹配绝对数值，从而提升蒸馏效果[15]。

5.2. 面向视觉 Transformer 的蒸馏损失

ViT等基于Transformer的视觉模型在图像分类、目标检测等任务中展现出强大性能，但其参数量和计算复杂度也随之增加。针对ViT的知识蒸馏需要考虑其独特的自注意力机制和patch embedding结构。当前研究主要聚焦于注意力关系的迁移和多层次特征对齐。

注意力关系蒸馏是ViT蒸馏的核心策略之一。自注意力机制通过计算token间的相似度矩阵捕获全局依赖关系，这种关系蕴含了丰富的结构化知识。DeiT [16]提出通过蒸馏token机制实现知识迁移，其损失函数包含标准分类损失和蒸馏损失：

$$\mathcal{L}_{\mathrm{DeiT}}=\mathcal{L}_{\mathrm{CE}}\left(y,\boldsymbol{z}_s^{\mathrm{cls}}\right)+\tau^2\cdot D_{\mathrm{KL}}\left(\sigma\left(\boldsymbol{z}_t/\tau\right)\|\,\sigma\left(\boldsymbol{z}_s^{\mathrm{dist}}/\tau\right)\right) \tag{12}$$

其中 $\boldsymbol{z}_s^{\mathrm{cls}}$ 和 $\boldsymbol{z}_s^{\mathrm{dist}}$ 分别为学生模型的分类token和蒸馏token输出，$\sigma(\cdot)$ 为softmax函数。

进一步地，注意力图蒸馏直接约束学生模型学习教师的注意力模式。对于多头注意力机制，可以最小化教师与学生注意力图之间的差异：

$$\mathcal{L}_{\mathrm{ATT}}=\frac{1}{LH}\sum_{l=1}^{L}\sum_{h=1}^{H}\left\|\boldsymbol{A}_t^{l,h}-\boldsymbol{A}_s^{l,h}\right\|_2^2 \tag{13}$$

其中 L 为层数，H 为注意力头数，$\boldsymbol{A}^{l,h}$ 表示第 l 层第 h 个头的注意力矩阵。这种方法使学生模型在token关系建模上更接近教师，从而提升表征能力[17]。

5.3. 面向大语言模型的蒸馏损失

大语言模型(LLM)的蒸馏面临着独特挑战：模型规模巨大、生成任务复杂、推理能力难以量化。传统的logit蒸馏在LLM场景下需要适应序列生成的特点，而中间层知识的选择性迁移成为平衡效率与效果的关键。

序列级知识蒸馏是LLM蒸馏的基础方法。与分类任务不同，语言生成需要在每个时间步进行预测，因此蒸馏损失需要聚合整个序列的信息。给定输入序列 $\boldsymbol{x}$，教师模型和学生模型在每个位置 t 产生的词汇表分布分别为 $p_t^{(t)}$ 和 $p_s^{(t)}$，序列级蒸馏损失定义为：

$$\mathcal{L}_{\text{seq}} = \frac{1}{T}\sum_{t=1}^{T} D_{\text{KL}}\left(p_t^{(t)} \,\|\, p_s^{(t)}\right) \tag{14}$$

其中 T 为序列长度。为了增强学生模型的生成能力，可以结合on-policy蒸馏，即让学生模型从自身生成的序列中学习[18]。

隐状态对齐则关注模型内部表征的迁移。由于LLM层数众多(如GPT-3有96层)，逐层对齐计算成本过高。实践中常采用间隔采样策略，选择关键层进行对齐。假设教师模型有 L_t 层，学生模型有 L_s 层，可以通过线性映射将学生第 l 层隐状态投影到教师相应层的空间：

$$\mathcal{L}_{\text{hidden}} = \sum_{l\in\mathcal{S}} \left\| \boldsymbol{W}_l \boldsymbol{H}_s^{(l)} - \boldsymbol{H}_t^{\left(f(l)\right)} \right\|_2^2 \tag{15}$$

其中 $\mathcal{S}$ 为选定的学生层索引集合，$f(l)$ 为学生层到教师层的映射函数，$\boldsymbol{W}_l$ 为可学习的投影矩阵。这种方法在保持计算可行性的同时，有效传递了教师模型的中间表征知识[19]。

6. 性能基准对比

为了直观展示不同损失函数在相同实验设定下的效果，本节提供了一个统一的性能基准对比。该基准选取了图像分类领域广泛使用的 CIFAR-100 数据集，并固定师生架构为：教师模型为 ResNet32x4，学生模型为 ShuffleNetV2。表中“学生(蒸馏)”列括号内的数值代表相较于学生基线模型的准确率提升(ΔAcc)。

Table 2. Performance comparison of knowledge distillation loss functions on CIFAR-100 database (Teacher: ResNet32x4, Student: ShuffleNetV2)
表 2. CIFAR-100 数据集上知识蒸馏损失函数性能对比(教师：ResNet32x4，学生：ShuffleNetV2)

Brown 方法	知识类型(匹配方式)	教师 Acc@1 (%)	学生基线 Acc@1 (%)	学生(蒸馏) Acc@1 (%)
FitNet [5]	特征匹配(值)	79.42	71.82	73.54 (+1.72)
KD [4]	Logit 匹配(熵)	79.42	71.82	74.45 (+2.63)
AT [6]	注意力图(值)	79.42	71.82	72.73 (+0.91)
SPKD [7]	相似性矩阵(值)	79.42	71.82	74.56 (+2.74)
CRD [13]	特征对比(值)	79.42	71.82	75.65 (+3.83)
DKD [14]	Logit 解耦(熵)	79.42	71.82	77.07 (+5.25)
DKD + logit 标准化[15]	Logit (熵)	79.42	71.82	77.37 (+5.55)
NormKD [20]	Logit 归一化(熵)	79.42	71.82	78.07 (+6.25)
CRLD [21]	相似性 + Logit (值 + 熵)	79.42	71.82	78.27 (+6.45)

结果分析

从表 2 的对比数据中可以观察到几个明显趋势：

1) 知识类型的演进带来性能提升。早期的特征匹配方法(如 FitNet，AT)提升幅度相对有限(+0.91%到+1.72%)。而迁移更高阶的结构化知识，如样本间相似性(SPKD)或通过对比学习构建的关系(CRD)，能带来更显著的增益(+2.74%到+3.83%)。

2) Logit 蒸馏的复兴与创新。经典的 KD 方法已经表现出色(+2.63%)，而通过对 Logit 知识进行更精细的设计——无论是将其解耦为目标类与非目标类(DKD, +5.25%)，还是进行归一化处理以聚焦于类间关系(NormKD, +6.25%)——都取得了突破性的性能，证明了 Logit 中蕴含的知识远未被充分挖掘。

3) 知识融合是前沿方向。表现最好的方法(CRLD, +6.45%)结合了多种知识类型(相似性 + Logit)，这表明融合不同维度、互补的知识是推动性能边界的关键。

该基准表明，损失函数的设计从简单的特征值模仿，发展到对结构化关系和精细化 Logit 知识的迁移，是性能得以持续提升的核心驱动力。

7. 总结与展望

本文系统回顾了知识蒸馏中损失函数的发展历程，从最初的基于输出响应匹配，到基于中间层特征匹配，再到目前前沿的基于关系与结构化知识匹配。这一演进过程体现了研究者对神经网络中“知识”本质的理解不断深化：从最终的决策结果，到中间层的抽象特征，再到特征或样本间的内在关系。

尽管损失函数的形式随着知识的载体而不断丰富，但其数学核心始终根植于两类基础的度量思想：基于“熵”的分布差异度量(如 KL 散度)与基于“值”的点对点差异度量(如均方误差 MSE)。领域创新的关键驱动力，更多在于探索“匹配什么”(What to Match)——即定义和提取何种形式的知识，而非发明“用什么匹配”(What to Measure)的基础数学工具。研究通过将“熵”与“值”这两类核心思想，创造性地应用于从软化输出、中间特征到高阶关系矩阵等不断深化的知识载体上，推动了整个领域的演进。

这一从“具体数值”到“抽象关系”的演进脉络，可以通过示意图(图 1)来清晰展示。它直观地描绘了知识定义如何从点(输出值)到线(特征图)，再到面(关系结构)的扩展过程，从根本上揭示了知识蒸馏技术发展的内在统一逻辑。

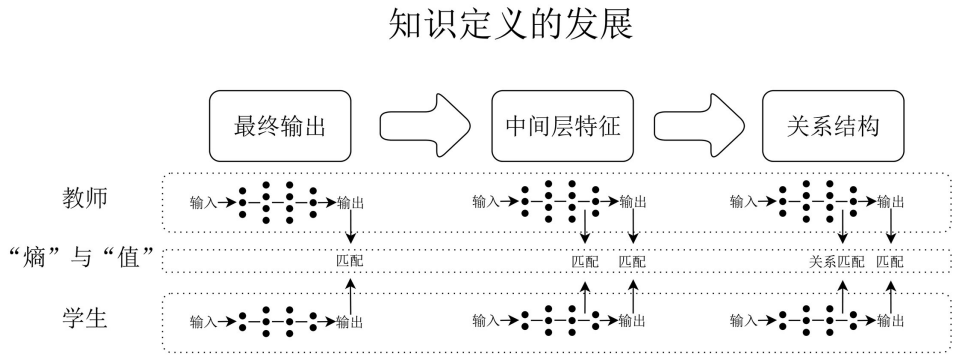

Figure 1. Evolution and context of knowledge definition in knowledge distillation
图 1. 知识蒸馏中知识定义的演进脉络示意图

当前挑战与未来展望

尽管取得了显著进展，知识蒸馏损失函数的研究仍面临一些挑战，并呈现出以下趋势：**1) 自适应与组合损失：**不同类型的损失函数迁移不同维度的知识，各有优劣。未来的一个重要方向是研究如何根据任务、数据、网络结构自适应地选择和组合多种损失函数，甚至通过元学习等方式自动学习损失权重。**2) 面向特定任务与结构的定制化损失：**在语音识别(如端到端模型)、目标检测、语义分割等特定任务中，网络结构和目标具有特殊性。设计针对这些场景的定制化损失函数，例如更好地对齐序列输出或空间特征图，具有重要的应用价值。**3) 理论分析的深化：**目前对于不同损失函数为何有效、它们迁移了何种具体知识、以及其与学生网络泛化能力之间关系的理论分析仍然不足。更坚实的理论基础将指导更有效的损失函数设计。**4) 超大规模模型的高效蒸馏：**面对参数规模巨大的预训练教师模型，如何设计高效的损失函数，在可承受的计算开销下实现有效的知识提取，是一个亟待解决的现实问题。**5) 跨模态与无标签蒸馏：**在教师和学生模型模态不同，或缺乏有标签数据的情况下，如何设计损失函数以实现知识迁移，

是一个具有挑战性的前沿课题。

总之，损失函数作为知识蒸馏的灵魂，其设计仍是该领域充满活力的研究方向。随着对神经网络表示学习的理解不断加深，以及新应用场景的不断涌现，未来必将出现更多高效、优雅的损失函数设计，持续推动知识蒸馏技术的发展与应用。

参考文献

[1] Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., *et al*. (2025) DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning. *Nature*, **645**, 633-638. https://doi.org/10.1038/s41586-025-09422-z

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2021) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. 9*th International Conference on Learning Representations*, *ICLR* 2021, 3-7 May 2021, 611-631. https://openreview.net/forum?id=YicbFdNTTy

[3] Buciluǎ, C., Caruana, R. and Niculescu-Mizil, A. (2006) Model Compression. Proceedings of the 12*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, 20-23 August 2006, 535-541. https://doi.org/10.1145/1150402.1150464

[4] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network.

[5] Romero, A., Ballas, N., Kahou, S.E., *et al*. (2015) FitNets: Hints for Thin Deep Nets. https://arxiv.org/abs/1412.6550

[6] Zagoruyko, S. and Komodakis, N. (2017) Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. 5*th International Conference on Learning Representations*, *ICLR* 2017, Toulon, 24-26 April 2017, 1489-1501. https://openreview.net/forum?id=Sks9_ajex

[7] Tung, F. and Mori, G. (2019) Similarity-Preserving Knowledge Distillation. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 1365-1374. https://doi.org/10.1109/iccv.2019.00145

[8] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 7132-7141. https://doi.org/10.1109/cvpr.2018.00745

[9] Zhou, Z., Zhuge, C., Guan, X., *et al*. (2020) Channel Distillation: Channel-Wise Attention for Knowledge Distillation. https://arxiv.org/abs/2006.01683

[10] Yim, J., Joo, D., Bae, J. and Kim, J. (2017) A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 7130-7138. https://doi.org/10.1109/cvpr.2017.754

[11] Peng, B., Jin, X., Li, D., Zhou, S., Wu, Y., Liu, J., *et al*. (2019) Correlation Congruence for Knowledge Distillation. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 5006-5015. https://doi.org/10.1109/iccv.2019.00511

[12] Wang, K., Vicol, P., Lucas, J., *et al*. (2018) Adversarial Distillation of Bayesian Neural Network Posteriors. https://arxiv.org/abs/1806.10317

[13] Tian, Y., Krishnan, D. and Isola, P. (2020) Contrastive Representation Distillation. ICLR. https://openreview.net/forum?id=SkgpBJrtvS

[14] Zhao, B., Cui, Q., Song, R., Qiu, Y. and Liang, J. (2022) Decoupled Knowledge Distillation. 2022 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 11943-11952. https://doi.org/10.1109/cvpr52688.2022.01165

[15] Sun, S., Ren, W., Li, J., Wang, R. and Cao, X. (2024) Logit Standardization in Knowledge Distillation. 2024 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 15731-15740. https://doi.org/10.1109/cvpr52733.2024.01489

[16] Touvron, H., Cord, M., Douze, M., *et al*. (2020) Training Data-Efficient Image Transformers & Distillation through Attention. https://arxiv.org/abs/2012.12877

[17] Yang, Z., Li, Z., Zeng, A., Li, Z., Yuan, C. and Li, Y. (2024) ViTKD: Feature-Based Knowledge Distillation for Vision Transformers. 2024 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Seattle, 16-22 June 2024, 1379-1388. https://doi.org/10.1109/cvprw63382.2024.00145

[18] Agarwal, R., Vieillard, N., Zhou, Y., *et al*. (2024) On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. *The Twelfth International Conference on Learning Representations*, *ICLR* 2024, Vienna, 7-11 May 2024, 9249-9266. https://openreview.net/forum?id=3zKtaqxLhW

[19] Wang, W., Wei, F., Dong, L., *et al*. (2020) MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. https://arxiv.org/abs/2002.10957

[20] Chi, Z., Zheng, T., Li, H., *et al*. (2023) NormKD: Normalized Logits for Knowledge Distillation. https://arxiv.org/abs/2308.00520

[21] Zhang, W., Liu, D., Cai, W. and Ma, C. (2024) Cross-View Consistency Regularisation for Knowledge Distillation. *Proceedings of the* 32*nd ACM International Conference on Multimedia*, *MM* 2024, Melbourne, 28 October-1 November 2024, 2011-2020. https://doi.org/10.1145/3664647.3681206