arXiv:2601.00451v1 [cs.LG] 1 Jan 2026

可控概念瓶颈模型

Hongbin Lin*、Chenyang Ren*、Juangui Xu*、Zhengyu Hu、Cheng-Long Wang、YaoShu、Hui Xiong，*IEEE* 会士，Jingfeng Zhang，*IEEE* 会员，Di Wang 和 Lijie Hu†

摘要—概念瓶颈模型（CBMs）凭借其通过人类可理解的概念层阐释预测过程的能力，受到了广泛关注。然而，以往大多数研究聚焦于静态场景，并假设数据和概念均固定且干净。在现实世界应用中，已部署的模型需要持续维护：我们经常需要删除错误或敏感数据（机器遗忘）、纠正错误标注的概念，或纳入新获取的样本（增量学习），以适应不断变化的环境。因此，如何在无需从头重新训练的情况下，高效地构建可编辑的概念瓶颈模型，仍是一项重大挑战，尤其是在大规模应用中。为应对这些挑战，我们提出了可控概念瓶颈模型（CCBMs）。具体而言，CCBMs 支持三种粒度的模型编辑：概念标签级、概念层级和数据级；其中数据级编辑涵盖数据删除和数据添加。CCBMs 利用影响函数推导出具有严格数学依据的闭式近似，从而无需重新训练。实验结果表明，我们的 CCBMs 高效且适应性强，证实了其在实现动态、可信赖的概念瓶颈模型方面的实用价值。

索引词—概念瓶颈模型、可解释人工智能、模型编辑、影响函数、机器遗忘、增量学习。

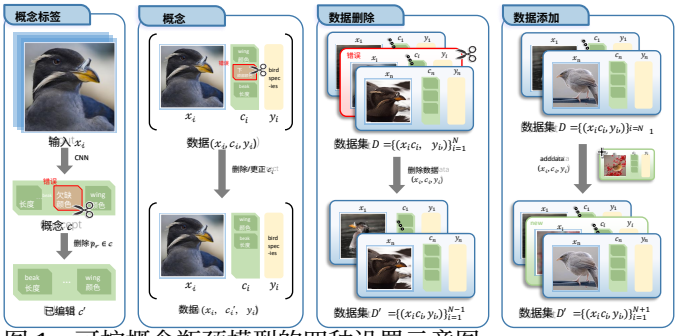


图 1：可控概念瓶颈模型的四种设置示意图。

1 引言

THE 深度学习在包括医疗保健 [1]，[2] 和金融 [3]，在内的高风险领域中迅速普及，同时也引发了人们对这些系统不透明性的日益担忧。尽管大语言模型 [4]–[8] 和多模态架构 [9]–[11] 取得了卓越的性能，但其复杂且非线性的决策过程通常仍令最终用户难以理解。为弥合这一信任鸿沟，可解释人工智能（XAI） [12]–[14] 已成为研究的重点。在各种可解释人工智能范式中，概念瓶颈模型（概念瓶颈模型） [15] 通过明确构建预测过程而脱颖而出：输入首先被映射到人类可理解的概念，随后作为最终分类的唯一依据。这种架构约束确保模型的推理过程透明，并与人类语义知识保持一致。

尽管概念瓶颈模型具有可解释性优势，但其实际部署仍面临重大障碍。以往文献主要关注两个方面：通过半监督或无监督学习 [16]–[18]，降低概念标注的高昂成本，以及缩小概念瓶颈模型与标准黑箱模型之间的性能差距 [17], [19], [20]。然而，这些研究通常将训练完成后的概念瓶颈模型视为静态实体。这一假设与现实世界应用的动态特性相矛盾：模型必须持续适应，以纠正错误、扩充知识库，并响应不断变化的数据治理政策。

在本研究中，我们认为**可控性**——即在部署后无需进行昂贵的重新训练，便能精确修改模型行为的能力——与可解释性同等重要。我们指出了三种需要这种控制的不同粒度，因此必须从静态概念瓶颈模型转向动态框架（见图 1）：

- 概念标签级可控性：人工标注容易出错。当专家反馈指出训练数据中的特定概念被错误标记时，相关方需要能够修正这些特定于标签的错误，以纠正模型的逻辑，同时保留样本其余部分所包含的有效信息。

- 概念级可控性：“有效知识”的定义会随时间演变。医学研究可能会发现需要添加到现有模型概念集中的新风险因素（例如肺癌的某些特定合并症）。反之，过去依赖的概念也可能被认为是虚假的或带有歧视性的（例如疫情预测中的人口统计偏差 [21]），因此必须立即移除。可控概念瓶颈模型必须支持这种本体的实时演化。

数据级可控性： 我们进一步将数据级控制分为两个双向流：

• * *H. Lin*、*C. Ren* 和 *J. Xu* 对本研究贡献相同。
• † *L. Hu* 为通讯作者。
• 稿件收到日期： *[Month] [Day], [Year]*；修订日期： *[Month] [Day], [Year]*。

0000–0000/00$00.00 © 2021年 IEEE

- 数据遗忘（移除）： 为遵守隐私法规（例如 GDPR）或丢弃投毒数据样本，模型必须能够从其参数中彻底消除特定数据点的影响。

- 数据添加（增量学习）：相反，随着新数据的不断出现，模型应能无缝整合这些新信息，以增强鲁棒性或弥补代表性不足，同时避免全面重新训练模型所带来的停机时间。

通过简单的重新训练来满足这些需求，在计算上代价高昂，尤其是对于大规模骨干网络。为克服这一瓶颈，我们提出了可控概念瓶颈模型（CCBMs）。与传统方法不同，CCBMs 能够根据上述变化，高效且有数学依据地更新模型参数。我们基于影响函数推导出闭式近似 [22], [23], ，并将其应用扩展到概念瓶颈模型的复合架构——这是一个其效用尚未得到充分探索的领域。

我们对该领域的贡献主要有以下四点：

- 我们将概念瓶颈模型中的事后可控性问题概念化，并定义了一个全面的分类体系，涵盖概念标签校正、概念集演化以及双向数据流（添加和移除）。
- 我们引入了CCBM框架，该框架利用影响函数来实现精确且低成本的模型更新，从而有效避免从头重新训练的需求。
- 我们提出了一种融合特征值校正克罗内克分解近似曲率（EK-FAC）的加速算法，使我们的方法能够扩展到高维参数空间。
- 大量实验证实，CCBMs在效用保持与计算效率之间实现了更优的权衡，展现出其作为可持续且可信人工智能基础框架的潜力。

2 相关工作

概念瓶颈模型。 CBM [15] 是用于图像分类和视觉推理的一种创新性深度学习方法。它在深度神经网络中引入概念瓶颈层，通过学习特定概念来增强模型泛化能力和可解释性。然而，CBM 面临两项主要挑战：其性能通常落后于不含概念瓶颈层的原始模型，这归因于从原始数据中提取到瓶颈特征的信息不完整。此外，CBM 依赖于费时费力的数据集标注。研究人员针对这些挑战探索了相应的解决方案。 [24]将 CBM 扩展到交互式预测设置中，引入交互策略来确定需要标注哪些概念，从而改进最终预测结果。 [16] 针对 CBM 的局限性，提出了一种名为无标签概念瓶颈模型的新框架。这种创新方法无需带标签的概念数据，即可将任意神经网络转换为具有可解释性的 CBM，同时保持较高的准确率。事后概念瓶颈模型 [17] 可以应用于各种神经网络而不影响模型性能，同时保留可解释性方面的优势。CBM 在图像领域的相关工作还包括 [25], [26], [27], [28] 和 [29]的研究。尽管已有许多关于 CBM 的研究，但我们首次通过影响函数探究概念之间的交互影响。我们的研究旨在通过在 CBM 中利用影响函数来弥合这一空白，从而解析概念模型之间的交互，并为概念编辑提供自适应解决方案。更多相关工作请参阅附录16。

3 预备知识

概念瓶颈模型。在本文中，我们考虑原始CBM，并采用[15]所使用的符号。我们考虑一个分类任务，其概念集表示为 $\{p_1, \cdots, p_k\}$，其中每个 p_i 都是由专家或大语言模型给出的概念；训练数据集表示为 $\mathcal{D} = \{z_i\}_{i=1}^n$，其中 $z_i = (x_i, y_i, c_i)$。这里，对于 $i \in [n]$，$x_i \in \mathbb{R}^{d_i}$ 表示输入特征向量，$y_i \in \mathbb{R}^{d_o}$ 表示标签（其中 d_o 对应类别数），而 $c_i = (c_i^1, \cdots, c_i^k) \in \mathbb{R}^k$ 表示概念向量。在此背景下，c_i^j 表示第i个数据的概念 p_j 的标签。在CBM中，我们的目标是学习两种表示：一种称为概念预测器，将输入空间转换为概念空间，表示为 g：$\mathbb{R}_i^d \to \mathbb{R}^k$；另一种称为标签预测器，将概念空间映射到预测空间，表示为 f：$\mathbb{R}^k \to \mathbb{R}^{d_o}$。通常，这里的映射 f 是线性的。对于每个训练样本 $z_i = (x_i, y_i, c_i)$，我们考虑两个经验损失函数：概念预测器$\hat{}$ $\hat{}g$ 和标签预测器 f:

$$\hat{g} = \arg\min_g \sum_{i=1}^{n} \sum_{j=1}^{k} g^j(x_i)^\top \log(c_i^j), \tag{1}$$

其中，$g^j(*)$ 是预测的第 j 个概念。为简洁起见，将损失函数记为 $L_C(g(x_i)\ c_i) = \sum_{j=1}^k L_C^j(g(x_i)\ c_i)$，对应数据 (x_i, c_i)。得到概念预测器 $\hat{}g$ 后，标签预测器定义如下：

$$\hat{f} = \arg\min_f \sum_{i=1}^{n} L_Y\big(f(\hat{g}(x_i)), y_i\big), \tag{2}$$

其中 L_Y 表示交叉熵损失，与（1）类似。概念瓶颈模型通过双重精确预测可解释概念向量 $\hat{c} = \hat{}g(x)$（匹配概念 c）和最终$\hat{}$输出 $\hat{}y = f(\hat{c})$（匹配标签 y），确保通过明确的概念中介实现透明推理。此外，本文主要关注标签预测器 ff 为线性变换的情形，因为这有助于提升可解释性，便于追踪概念与标签之间的关系。有关所用符号的详细信息，请参阅附录中的符号表**??**。

影响函数。 影响函数衡量估计量对样本中单个$\hat{}$点取值的依赖程度。考虑一个神经网络 $\theta = \arg\min_\theta \sum_{i=1}^n \ell(z_i; \theta)$，其损失函数为 ℓ，数据集为$D = \{z_i\}_{i=1}^n$。如果从训练数据集中移除 z_m，$\hat{}$参数将变为 $\theta_{-z_m} = \arg\min_\theta \sum_{i \neq m} \ell(z_i; \theta)$。

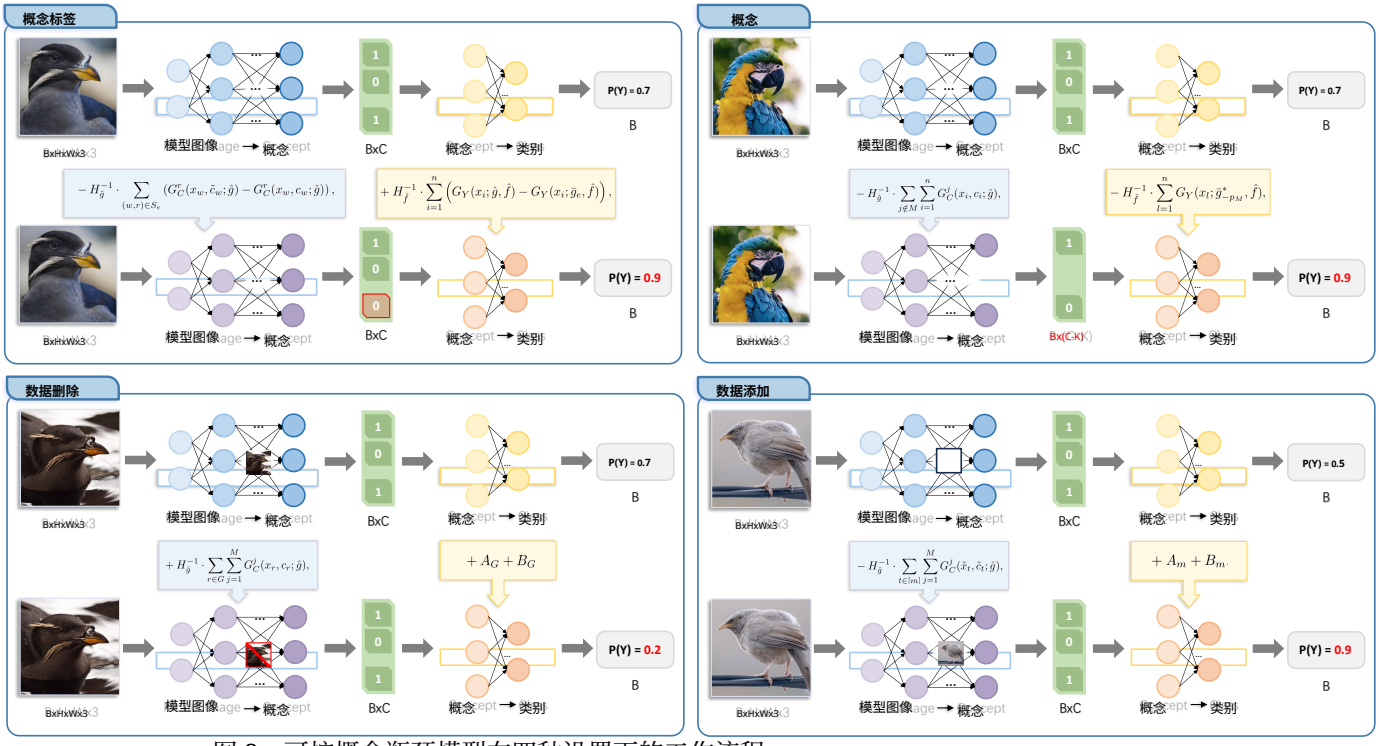


图 2：可控概念瓶颈模型在四种设置下的工作流程。

影响函数通过定义一系列由 ϵ 参数化的模型，提供高效的模型近似：$\hat{}\theta_{\epsilon,-z_m} = \text{argmin}\sum_{i=1}^{n} \ell(z_i;\theta) + \epsilon\ell(z_m;\theta)$。对对应于argmin 过程的目标函数梯度进行一阶泰勒展开后，影响函数定义如下：

$$\mathcal{I}_{\hat{\theta}}(z_m) \triangleq \left.\frac{\mathrm{d}\hat{\theta}_{\epsilon,-z_m}}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1}\cdot\nabla_\theta\ell(z_m;\hat{\theta}),$$

其中，$H^{-1}\hat{}_{\theta} = \nabla_\theta^2\sum_{i=1}^{n}\ell(z_i;\hat{}\theta)$ 是海森矩阵。当损失函数 ℓ 关于 θ 二次可微且强凸时，海森矩阵 $H_{\hat{\theta}}$ 为正定，因此影响函数有明确定义。对于非凸损失函数，[30] 提出用 $\hat{} H = G_{\hat{\theta}} + \delta I$ 替换海森矩阵 $H_{\hat{\theta}}$，其中 $G_{\hat{\theta}}$ 是费 Fisher 信息矩阵，定义为 $\sum_{i=1}^{n}\nabla_\theta\ell(z_i;\theta)^{\mathrm{T}}\nabla_\theta\ell(z_i;\theta)$，而 δ 是用于确保 H 正定的阻尼$\hat{}$ 项。我们可以采用特征值校正的 Kronecker 因子近似曲率（EK-FAC）方法，进一步加快计算。更多详情请参见附录9。

4 可控概念瓶颈模-型

本节将介绍我们针对引言中提到的三种设置提出的 CCBM，并利用影响函数进行分析。具体而言，在概念标签层级，我们计算一组数据样本中各个概念标签的影响；在概念层级，我们计算多个概念的影响；在数据级别，我们计算多个样本的影响。图2 展示了我们的 CCBM 在相应设置下的工作流程。

4.1 概念标签级可控概念瓶颈模型

在许多情况下，某些数据样本对特定概念包含错误标注，但其他信息仍然具有价值。这在医学影像等领域尤为常见，因为获取数据通常成本高昂且耗时。在这种情况下，人们通常会更正错误的概念标注，而不是将整个数据从数据集中移除。在此情境下，估计重新训练后的模型参数至关重要。我们将此情境称为概念标签层级可控概念瓶颈模型。

从数学上讲，我们有一组错误数据 D_e 及其关联索引集 $S_e \subseteq [n]\times[k]$，使得对于每个$(w,\ r) \in S_e$，$(x_w,\ y_w,\ c_w) \in D_e$，其中 c_w^r 被错误标记，而$\tilde{c}_w^r$ 是修正后的概念标签。我们的目标是估计重新训练的概念瓶颈模型。重新训练的概念预测器和标签预测器表示如下：

$$\begin{aligned}\hat{g}_e = \arg\min_g &\sum_{(i,j)\notin S_e} L_C^j\left(g(x_i), c_i\right) \\ &+ \sum_{(i,j)\in S_e} L_C^j\left(g(x_i), \tilde{c}_i\right),\end{aligned} \tag{3}$$

$$\hat{f}_e = \arg\min_f \sum_{i=1}^{n} L_Y\left(f\left(\hat{g}_e\left(x_i\right)\right), y_i\right). \tag{4}$$

对于简单的神经网络，我们可以直接使用影响函数方法来估计重新训练后的模型。然而，对于概念瓶颈模型架构，如果我们干预真实概念，概念预测器$\hat{}g$ 会相应地波动至$\hat{}g_e$。注意，标签预测器的输入数据来自概念预测器的输出，而这一输出也会发生变化。因此，我们需要采用两阶段编辑方法。在此，我们分别考虑（3）和（4）的影响函数。我们首先将概念预测器$\hat{}^{-}$从$\hat{}g$ 编辑为 $\bar{g}_e$，然后基于我们的

近似概念预测器，将{v18}编辑为{v19}。首先，我们给出以下定义：

定义 4.1. 定义第 j 个概念预测器以及第 i 个数据点 x_i 的标签预测器的梯度如下：

$$G_C^j(x_i, c_i; g) \triangleq \nabla_g L_C^j\left(g(x_i), c_i\right),$$
$$G_Y(x_i; g, f) \triangleq \nabla_f L_Y\left(f(g(x_i)), y_i\right).$$

定理4.2. 重新训练的概念预测器 $\hat{}g_e$ 由 *(3)* 定义，可近似为 $\bar{g}_e$，定义如下：

$$\hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{(w,r)\in S_e} \left(G_C^r(x_w, \tilde{c}_w; \hat{g}) - G_C^r(x_w, c_w; \hat{g})\right),$$

其中 $H_{\hat{g}} = \nabla_{\hat{g}} \sum_{i,\ j} G_C^j(x_i,\ \ c_i;\ \ \hat{}g)$ 是关于损失函数相对于 $\hat{}g$的海森矩阵。

定理4.3. 重新训练的标签预测器 f_e 由 (4) 可以近似为 f_e，定义如下：

$$\hat{f} + H_{\hat{f}}^{-1} \cdot \sum_{i=1}^{n} \Big(G_Y(x_i; \hat{g}, \hat{f}) - G_Y(x_i; \bar{g}_e, \hat{f})\Big),$$

其中 $H_{\hat{}f} = \nabla_{\hat{}f} \sum_{i=1}^{n} G_Y(x_i;\ \ \hat{}g,\ \ \hat{}f)$ 是海森矩阵，且 $\bar{g}_e$ 由定理 4.2给出。

与测试时干预的区别。 在概念瓶颈模型中进行干预的能力，使人类用户能够在预测过程中与模型交互。例如，医学专家可以直接替换错误预测的概念值$\hat{c}$，并观察其对最终预测$\hat{}y$的影响。然而，概念预测器存在的根本缺陷仍未得到解决，这意味着在新的测试数据上应用时，类似的错误可能依然存在。相比之下，在可控概念瓶颈模型框架下，不仅可以进行测试时干预，还可以根据反复产生错误的测试数据进一步优化概念瓶颈模型的概念预测器。我们的 CCBM 方法会将纠正后的测试数据纳入训练数据集，而无需完全重新训练。此方法将纠正过程从数据级别扩展到了模型级别。

4.2 概念层级可控概念瓶颈模型

在这种情况下，由于归因错误或存在伪概念，一组概念被移除，这称为概念级编辑。[1]具体而言，对于概念集，将错误的概念索引集记为 $M \subset [k]$，我们的目标是删除所有训练样本中的这些概念标签。我们旨在研究更新训练数据中的概念集对模型预测的影响。值得注意的是，与上述概念标签情形相比，重新训练的概念预测器的输出（标签预测器的输入）维度会发生变化。如果从数据集中删除 t 个概念，那么 g 就变为 g'： $\mathbb{R}^{d_i} \to \mathbb{R}^{k-t}$，而 f 就变为 f'： $\mathbb{R}^{k-t} \to \mathbb{R}^{d_o}$。更具体地说，如果使用修订后的数据集重新训练概念瓶颈模型，相应的概念预测器将变为：

$$\hat{g}_{-p_M} = \arg\min_{g'} \sum_{j\notin M} \sum_{i=1}^{n} L_C^j(g'(x_i), c_i). \tag{5}$$

1. 为方便起见，本文仅考虑概念删除；我们的方法可以直接扩展到概念插入。

参数在维度上的变化，使得基于影响函数的编辑难以应用于概念预测器。这是因为影响函数通过对给定样本对应的损失进行 ϵ 加权，并近似计算相对于原始基准的参数变化，来实现对预测器的编辑，因此无法处理参数维度的变化。

为克服这一挑战，我们的策略是对$\hat{}g_{-p_M}$进行一些变换，使其维度与$\hat{}g$相匹配，从而能够应用影响函数来编辑概念瓶颈模型。我们通过将$\hat{}g_{-p_M}$映射到$\hat{}g^*_{-p_M} \triangleq \mathrm{P}(\hat{}g_{-p_M})$来实现这一点。后者的参数数量与$\hat{}g$相同，并且对于所有 $j \in [d_i] - M$，其预测概念$\hat{}g^*_{-p_M}(j)$与$\hat{}g_{-p_M}(j)$相同。我们通过在$\hat{}g_{-p_M}$最后一层的矩阵第 r 行插入一个零行向量来实现这一效果，其中 $r \in M$。因此，我们可以看出，映射 P 是一一对应的。此外，假设$\hat{}g$的参数空间为 T，而$\hat{}g^*_{-p_M}$, T_0 的参数空间是 T 的子集。注意到，$\hat{}g^*_{-p_M}$是以下目标函数的最优模型：

$$\hat{g}^*_{-p_M} = \arg\min_{g'\in T_0} \sum_{j\notin M} \sum_{i=1}^{n} L_C^j(g'(x_i), c_i), \tag{6}$$

也就是说，它是在约束 T_0 下，针对剩余概念的概念预测器损失的最优模型。现在，我们可以应用影响函数来编辑$\hat{}g$，使其近似$\hat{}g^*_{-p_M}$，同时限制神经网络最后一层中由 M 索引的行的值为 0，记为 $\bar{g}^*_{-p_M}$。之后，我们从 $\bar{g}^*_{-p_M}$ 中移除最初为填补维度差异而插入的参数；由于编辑阶段施加的约束，这些参数始终为 0，从而近似得到真正编辑后的概念预测器 $\hat{}g_{-p_M}$。接下来，我们通过以下定理详细说明从 $\hat{}g$ 到$\hat{}g^*_{-p_M}$ 的编辑过程。

定理 4.4. 对于$\hat{}g_{-p_M}$中定义的重新训练后的概念预测器 (5)，我们将其映射为 $\hat{}g^*_{-p_M}$，如(6)所示。我们还可以将初始的$\hat{}g$ 编辑为$\hat{}g^*_{-p_M}$，定义如下：

$$\bar{g}^*_{-p_M} \triangleq \hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{j\notin M} \sum_{i=1}^{n} G_C^j(x_i, c_i; \hat{g}), \tag{7}$$

其中 $H_{\hat{g}} = \nabla_g \sum_{j\notin M} \sum_{i=1}^{n} G_C^j(x_i,\ \ c_i;\ \ \hat{}g)$。然后，通过移除映射阶段插入的所有零行，我们可以自然地近似 $\hat{}g_{-p_M} \approx \mathrm{P}^{-1}(\hat{}g^*_{-p_M})$。

对于训练的第二阶段，假设我们的目标是移除 p_r 中的概念 $r \in M$，且新的最优模型为$\hat{}f_{-p_M}$。我们将遇到与第一阶段相同的困难，即标签预测器的参数数量会发生变化。为解决这一问题，我们的关键观察是，在现有关于概念瓶颈模型的文献中，我们始终对标签预测器使用线性变换，这意味着输入中取值为0 的维度不会对最终预测产生贡献。利用这一性质，我们将更新后的预测器输入中的缺失值填充为0，即用$\hat{}g^*_{-p_M}$ 替换$\hat{}g_{-p_M}$，并令$\hat{}$ $f_{p_M=0}$ 定义为

$$\hat{f}_{p_M=0} = \arg\min_{f} \sum_{i=1}^{n} L_Y\left(f\left(\hat{g}^*_{-p_M}(x_i)\right), y_i\right). \tag{8}$$

总之，我们有以下引理：

引理4.5。 在概念瓶颈模型中，如果标签预测器对输入 c采用形式为 $f \cdot c$ 的线性ˆ变换，那么，对于每个$r \in M$，我们从c中移除第 r个概念，并将新的输入记为 c'；将第 r个概念设为0，并将新的输入记为c^0。那么，对于任意输入 c，都有ˆ $f_{-p_M} \cdot c' = \hat{f}_{p_M=0} \cdot c^0$ 。

ˆ引理4.5 表明，重新训练后的 f_{-p_M} 和ˆ$f_{p_M=0}$分别以ˆ $g_{-p_M}(x)$ 和ˆ$g^*_{-p_M}(x)$ 作为输入时，会产生相同的输出。因此，我们可以使用ˆ ˆ$f_{p_M=0}$ 作为编辑目标，以替代 f_{-p_M}。

ˆ**定理4.6.** 对于由(8)定义的修订版重新训练的标签预测器 $f_{p_M=0}$ˆ ¯，我们可以通过以下方程将初始标签预测器 f 编辑为 $f_{p_M=0}$ ˆ，以替代 $f_{p_M=0}$:

$$\hat{f}_{p_M=0} \approx \bar{f}_{p_M=0} \triangleq \hat{f} - H^{-1}_{\hat{f}} \cdot \sum_{l=1}^{n} G_Y(x_l; \bar{g}^*_{-p_M}, \hat{f}), \quad (9)$$

其中 $H_{\hat{f}} = \nabla_{\hat{f}} \sum_{i=1}^{n} G_Y(x_l; \bar{g}^*_{-p_M}, \hat{f})$ 是海森矩阵。¯删除 $f_{p_M=0}$中对应于 $r \in M$的第 r个维度后，我们可以将其映射到 f_{-p_M}，这就是概念层级下最终编辑后的标签ˆ预测器f_{-p_M} 的近似。

4.3 数据级可控概念瓶颈模型

本节将探讨数据级的可控概念瓶颈模型。我们确定了已部署模型生命周期中两个不同但同等关键的操作需求：数据遗忘（移除特定样本）和数据添加 （纳入新样本）。我们针对这两种情形分别推导了专门的近似方法。

4.3.1 数据遗忘（移除）

在这种情形下，主要目标是彻底消除特定数据样本对已训练概念瓶颈模型的影响。这一要求通常源于隐私合规需求（例如“被遗忘权”），或需要清除模型中的投毒或错误标注的训练样本。

从数学上来说，设 $\mathcal{D}$ 为原始数据集。我们确定一个待移除的样本子集，记为 $G \subset \mathcal{D}$，其中包含元素 $z_r = (x_r, y_r, c_r)$。目标是近似求得模型参数，使其效果如同在净化后的数据集 $\mathcal{D} \setminus G$ 上训练所得。重新训练的概念预测器定义如下：

$$\hat{g}_{-z_G} = \arg\min_{g} \sum_{j=1}^{k} \sum_{z_i \in \mathcal{D} \setminus G} L^j_C(g(x_i), c_i). \quad (10)$$

由于重新训练的计算成本高昂，我们采用影响函数来估计移除 G 的损失梯度所导致的参数变化。

定理 4.7。 给定要移除的子集 G，净化后的概念预测器 ˆ g_{-z_G} 定义于(10) 可近似为：

$$\hat{g}_{-z_G} \approx \bar{g}_{-z_G} \triangleq \hat{g} + H^{-1}_{\hat{g}} \cdot \sum_{z_r \in G} \sum_{j=1}^{M} \nabla_g L^j_C(x_r, c_r; \hat{g}), \quad (11)$$

其中， $H_{\hat{g}}$ 是概念总损失的海森矩阵。添加符号表示一种逆操作，即消除 G 中梯度的负面影响。

随后，标签预测器必须根据更新后的概念表示来定义其决策边界。重新训练的标签预测器对应于：

$$\hat{f}_{-z_G} = \arg\min_{f} \sum_{z_i \in \mathcal{D} \setminus G} L_Y\left(f(\hat{g}_{-z_G}(x_i)), y_i\right). \quad (12)$$

由于训练集大小和输入特征（更新后的概念）同时发生变化，直接估计难以实现。我们采用两阶段估计策略。首先，˜我们定义一个中间预测器 f_{-z_G}，它仅考虑数据删除，同时固定概念预测器：

$$\tilde{f}_{-z_G} = \arg\min_{f} \sum_{z_i \in \mathcal{D} \setminus G} L_Y(f(\hat{g}(x_i)), y_i). \quad (13)$$

定理 4.8。 遗忘设置下标签预测器的参数变化推导如下。首先，通过移除 G 的梯度影响来近似中间预测器：

$$\begin{aligned} \tilde{f}_{-z_G} &\approx \bar{f}^*_{-z_G} \\ &\triangleq \hat{f} + H^{-1}_{\hat{f}} \sum_{z_r \in G} \nabla_f L_Y(x_r; \hat{g}, \hat{f}) \\ &\triangleq \hat{f} + A_{unlearn}. \end{aligned}$$

其次，为了考虑从 ˆg 到 $\bar{g}_{-z_G}$ 的概念漂移，我们计算校正项 $B_{unlearn}$:

$$B_{unlearn} = -H^{-1}_{\bar{f}^*} \sum_{z_i \in \mathcal{D} \setminus G} \Big(\nabla_f L_Y(x_i; \bar{g}_{-z_G}, \bar{f}^*) - \nabla_f L_Y(x_i; \hat{g}, \bar{f}^*) \Big),$$

¯ ¯其中， f^* 表示 $f^*_{-z_G}$。最终机器遗忘后的标签预测器为 ¯ ˆ$f_{-z_G} = f + A_{unlearn} + B_{unlearn}$。

4.3.2 数据添加（增量适应）

与机器遗忘不同，现实世界中的系统通常必须适应持续不断的数据流。在这种数据添加场景中，通过吸收一批新的样本来扩展概念瓶颈模型，记为 $\mathcal{S}_{new} = \{(\tilde{x}_k, \tilde{y}_k, \tilde{c}_k)\}_{k=1}^{m}$，其中$m \ll n$。移除旨在撤销已有知识，而数据添加则旨在增强模型的能力，同时避免重新处理整个历史数据集所带来的高昂成本。这就形成了一个增量学习问题。

结合历史数据和新数据后，重新训练的目标概念预测器定义如下：

$$\hat{g}_{+\mathcal{S}_{new}} = \arg\min_{g} \left[\sum_{i=1}^{n} L_C(g(x_i), c_i) + \sum_{k=1}^{m} L_C(g(\tilde{x}_k), \tilde{c}_k) \right]. \quad (14)$$

我们不将其视为重新训练任务，而是将其表述为牛顿步更新，其中新数据提供一个梯度下降方向，用于修正现有参数。

定理 4.9。 使用 $\mathcal{S}_{new}$增广数据集时，更新后的概念预测器ˆ $g_{+\mathcal{S}_{new}}$ 可以通过以下方式高效近似：

$$\hat{g}_{+\mathcal{S}_{new}} \approx \bar{g}_{+\mathcal{S}_{new}} \triangleq \hat{g} - H^{-1}_{\hat{g}} \cdot \sum_{z_k \in \mathcal{S}_{new}} \sum_{j=1}^{M} \nabla_g L^j_C(\tilde{x}_k, \tilde{c}_k; \hat{g}). \quad (15)$$

这里，减号表示我们沿着新数据的梯度所指示的下降方向移动参数，以最小化增广损失。

备注（增量学习视角）。值得注意的是，定理 4.9实际上执行了二阶在线更新。通过利用从原始数据中计算得出的海森矩阵（可预先计算并存储） $H_{\hat{g}}$，我们将新样本的梯度 $\mathcal{S}_{new}$ 投影到现有流形的曲率上。这种方法使概念瓶颈模型能够“吸收”新知识，同时保持相对于先前概念的稳定性。

同样，标签预测器必须进行调整，以便为增广数据集定义决策边界。目标如下：

$$\hat{f}_{+\mathcal{S}_{new}} = \arg\min_{f} \Bigg[\sum_{i=1}^{n} L_Y(f(\hat{g}_{+\mathcal{S}_{new}}(x_i)), y_i) + \sum_{k=1}^{m} L_Y(f(\hat{g}_{+\mathcal{S}_{new}}(\tilde{x}_k)), \tilde{y}_k) \Bigg]. \tag{16}$$

˜我们定义中间预测器 $f_{+\mathcal{S}_{new}}$，它在原始概念映射 ˆg下吸收新数据。

定理 4.10。标签预测器的适应涉及两个不同的变化。首先，我们估计新样本对决策边界的影响：

$$\tilde{f}_{+\mathcal{S}_{new}} - \hat{f} \approx -H_{\hat{f}}^{-1} \sum_{z_k \in \mathcal{S}_{new}} \nabla_f L_Y(\tilde{x}_k; \hat{g}, \hat{f}) \triangleq A_{add}.$$

¯ ˆ设 $f^*_{+\mathcal{S}_{new}} = f + A_{add}$。其次，我们校正新数据导致的概念表示变化。校正项 B_{add}推导如下：

$$B_{add} = -H_{\bar{f}^*}^{-1} \sum_{z \in \mathcal{D} \cup \mathcal{S}_{new}} \Big(\nabla_f L_Y(z; \bar{g}_{+\mathcal{S}_{new}}, \bar{f}^*) - \nabla_f L_Y(z; \hat{g}, \bar{f}^*) \Big),$$

其中，$H_{\bar{f}^*}$ 是相对于中间预测器的海森矩阵。¯ ˆ最终更新后的标签预测器为 $f_{+\mathcal{S}_{new}} = f + A_{add} + B_{add}$。

4.3.3 通过 *EK-FAC* 加速。
如第 3 节所述，概念瓶颈模型中的损失函数是非凸的，这意味着我们所有定理中的海森矩阵都可能未定义。为解决这一问题，我们采用 EK-FAC 方法，将海森矩阵近似为 ˆ $H_\theta = G_\theta + \delta I$。其中，$G_\theta$ 表示模型 θ 的费 Fisher 信息矩阵，而 δ 是为确保正定性而引入的小阻尼项。有关将 EK-FAC 应用于概念瓶颈模型的详细信息，请参阅附录 9.1。此外，有关与三个级别对应的 EK-FAC 算法（其中数据级别包括添加和移除），请参阅附录中的算法 7-9；相应的原始（基于海森矩阵）版本分别见算法1-4。

理论界。我们针对所有三个层级，给出了重新训练与 CCBM 之间概念预测器的误差界；详情请参见附录 10.1，11.1 和12.3。我们从理论上证明，在某些情形下，利用经海森矩阵中的阻尼项 δ 正则化的近似误差是可以容忍的。

5 实验

本节将全面评估所提出的可控概念瓶颈模型（CCBMs）。我们从三个主要维度评估该框架：效用保持、编辑效率和可解释性。由于篇幅限制，更多实验细节和补充结果请参见附录 15。

5.1 实验设置

数据集。我们在四个不同的基准测试上评估我们的方法：*X*射线分级 *(OAI)* [31]，鸟类识别 *(CUB)* [32],大规模名人面部属性 *(CelebA)* [33], 以及*Derm7pt* [34]。

- **OAI：** 一项针对膝骨关节炎的多中心观察性研究，包含 36,369 个数据点。我们配置了 $n = 10$ 个概念，用于表征关键的骨关节炎指标（例如，关节间隙狭窄、骨赘）。我们按照 [15]，中的方案对数据进行预处理，使其符合标准的概念瓶颈模型设置。

- **CUB：** 该数据集包含 11,788 张鸟类图像，涵盖 200 个类别，并标注了 112 个二元属性（概念）。我们遵循 [15] 中的处理设置。具体而言，我们通过多数投票将实例级概念标注汇总为类别级概念：例如，如果某个类别（如乌鸦）中超过 50% 的样本具有某项特征（如黑色翅膀），则将该概念赋予整个类别。

- **CelebA：** 一个大规模面部属性数据集，包含 202,599 张图像。每张图像均标注了 40 个二元属性。按照 [20]，，我们指定 8 个属性作为分类目标（标签），其余 32 个属性作为瓶颈概念。

- **Derm7pt：** 一个用于黑色素瘤诊断的皮肤镜基准数据集，包含 1,011 张图像。该数据集标注了 7 项临床定义的皮肤镜标准（例如色素网络、条纹、蓝白色幕），用作概念；此外还包含一个二元诊断标签（黑色素瘤或非黑色素瘤）。按照 [35]，我们使用官方数据划分和标准归一化流程。

Ev评估指标。我们采用两个主要 指标：

1) *F1* 分数：通过平衡精确率和召回率来衡量模型效用。
2) 运行时间（*RT*）：以分钟为单位测量，用于评估完成模型更新所需的计算时间。

实现细节。实验在 Intel Xeon CPU 和 NVIDIA RTX 3090 GPU 上进行。对于概念标签级 编辑，我们随机选择 3% 的数据点，并为每个数据点翻转一个概念标签，重复 10 次。对于概念层级 编辑，我们随机移除 OAI 中的一个概念和 CUB 中的十个概念，并使用五个不同的随机种子重复此过程。对于数据级 编辑，我们评估机器遗忘和添加场景：（1）对于移除，从完整数据集中随机删除 3% 的训练数据，重复 10 次。（2）对于添加，我们在训练阶段开始时随机留出 10% 的训练数据子集，并在

表1：不同方法在三个数据集上的性能比较。

编辑级别	方法	OAI		CUB		CelebA		Derm7pt	
		F1 分数	运行时间（分钟）	F1 分数	运行时间（分钟）	F1 分数	运行时间（分钟）	F1 分数	运行时间（分钟）
概念标签	重新训练	0.8825±0.0054	297.77	0.7971±0.0066	85.56	0.3827±0.0272	304.71	0.7826±0.0059	33.79
	CBM-IF(我们的方法)	0.8650±0.0030	4.58	0.7710±0.0033	1.28	0.3597±0.0128	5.50	0.7689±0.0060	0.44
	CCBM（我们的方法）	**0.8809±0.0033**	**2.28**	**0.7918±0.0042**	**0.63**	**0.3831±0.0326**	**2.40**	0.7813±0.0054	0.23
概念	重新训练	0.8448±0.0191	258.84	0.7811±0.0047	87.21	0.3776±0.0350	355.85	0.7697±0.0065	36.11
	CBM-IF(我们的方法)	0.8248±0.0065	4.87	0.7580±0.0059	1.44	0.3642±0.0195	5.44	0.7421±0.0064	0.38
	CCBM（我们的方法）	**0.8411±0.0081**	**2.29**	**0.7794±0.0056**	**0.51**	**0.3768±0.0278**	**2.40**	0.7609±0.0068	0.30
数据删除	重新训练	0.8811±0.0065	319.37	0.7838±0.0051	86.20	0.3797±0.0375	325.62	0.7753±0.0053	35.47
	CBM-IF（我们的方法）	0.8477±0.0039	4.99	0.7625±0.0022	1.41	0.3546±0.0163	5.89	0.7648±0.0035	0.56
	CCBM（我们的方法）	**0.8799±0.0038**	**2.42**	**0.7851±0.0079**	**0.60**	**0.3754±0.0337**	**2.41**	0.7728±0.0072	0.37
数据添加	重新训练	0.8801±0.0058	323.83	0.7987±0.0071	87.93	0.3830±0.0285	331.42	0.7850±0.0068	36.39
	CBM-IF（我们的方法）	0.8479±0.0039	5.03	0.7635±0.0022	1.44	0.3558±0.0163	5.95	0.7581±0.0036	0.60
	CCBM（我们的方法）	**0.8806±0.0030**	**2.51**	**0.7967±0.0080**	**0.63**	**0.3817±0.0340**	**2.47**	0.7831±0.0078	0.43

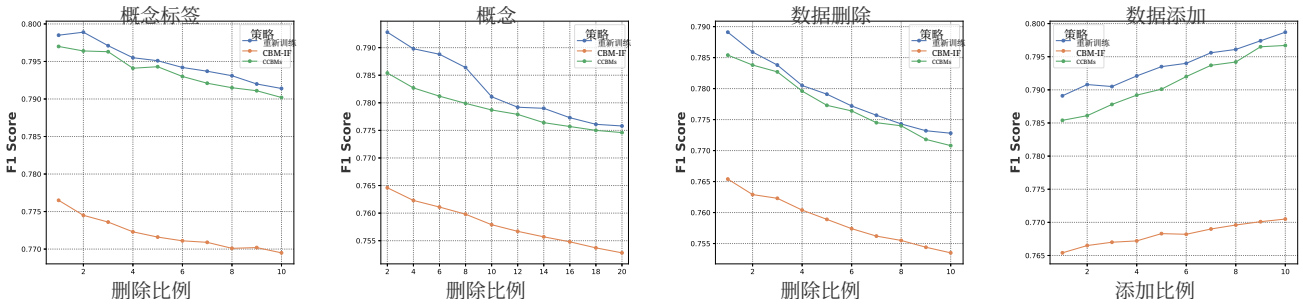


图3：编辑比例对CUB数据集上三种设置的影响。

首次训练阶段结束后将这些样本加回，以更新模型。为了测试模型在多样本编辑（图 3）中的鲁棒性，我们依次移除概念（从 2 到 20 个）并删除数据点（比例从 1% 到 10%）。

5.2 效用与编辑效率评估

与重新训练的比较。表 1 中总结的主要结果凸显了 CCBMs 相较于传统再训练和基线 CBM-IF方法的优势。其最显著的优势在于计算效率。例如，在 OAI数据集上，CCBMs 将更新时间从 297.77 分钟（重新训练）缩短至仅 2.36 分钟，提速超过 100×，同时 F1 分数（0.8808）与重新训练（0.8825）高度相当。这一趋势在 CUB数据集上同样明显：运行时间从 85.56 分钟降至 0.65 分钟，性能损失可以忽略不计（0.7971 与 0.7963）。与标准影响函数基线方法（CBM-IF）相比，借助 EK-FAC 的 CCBMs 不仅更新速度更快，F1 分数也始终更高。这一改进表明，EK-FAC 近似能够有效稳定海森矩阵求逆；在非凸情形下，海森矩阵通常会病态，因此这种近似能够提供更准确的更新方向。

编辑多个样本。 为了评估 CCBM 在大幅修改情况下的鲁棒性，我们通过编辑不同比例的样本和概念进行了压力测试。如图 3所示，CCBM 表现出卓越的稳定性。在数据级和概念标签级设置中，即使编辑多达 10% 的数据，CCBM 相较于重新训练的性能差距也微乎其微（差距 < 0.0025）。在概念层级，尽管随着移除的概念数量增加，差异略有扩大，CCBM 仍始终优于 CBM-IF。考虑到显著的时间节省（比重新训练快 3× 倍在这些批量处理场景中），CCBM 在编辑速度和模型效用之间实现了最佳权衡。

5.3 可解释性结果

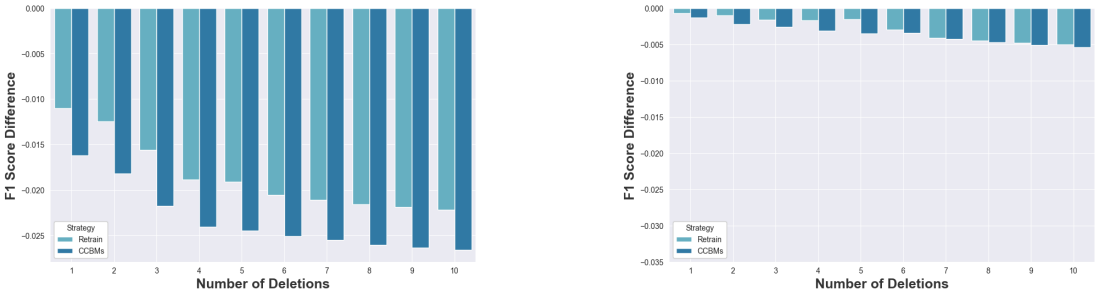
通过影响函数量化概念重要性。概念瓶颈模型的一项核心优势是其语义透明性。CCBMs 在此基础上进一步提供了严谨的数学框架，用于量化每个概念对最终预测的*贡献*。我们通过在 CUB数据集上进行“概念消融”实验，验证了这一能力。我们利用提出的影响函数，按照所有概念的影响力进行排序，并进行了两项对比测试：移除影响力最大的前 10 个概念，与移除影响力最小的后 10 个概念。

图 4a和图 4b（见图 **??**）展示了 F1 分数的变化趋势。我们观察到三个关键发现：

1）**判别能力：**移除影响力最大的概念（图 4a）会导致模型性能急剧且持续下降。仅移除 10 个概念后，F1 分数就下降了 0.025 以上（从 ≈ 0.780 降至 < 0.755）。这证实了影响函数能够正确识别出驱动模型决策的“承重”语义特征。

2) **噪声识别：** 相反，移除影响最小的概念（图 4b）仅会导致可忽略的性能变化（变化 < 0.005）。平坦轨迹表明，CCBM 能有效隔离冗余或无信息概念，防止模型过度依赖虚假特征。

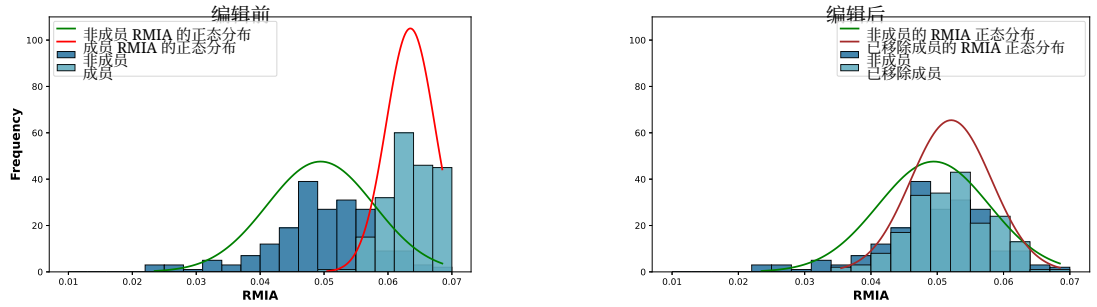
3）**与真实标签对齐：**最重要的是，CCBM（深蓝色柱）和重新训练（浅蓝色柱）的性能下降曲线非常相似。两种方法之间的差异



(a) 影响力最大的1–10个 con- 的结果
概念

(b) 移除影响最小的 1–10 个概念的结果

Fig. 图 4：CCBM 移除影响最大和最小的概念后的 F1 分数差异 .



(a) 编辑前的 RMIA 分数

(b) 编辑后的 RMIA 分数

图 5：移除前后的数据 RMIA 分数。

始终低于0。005。这种高度保真的表现表明，基于影响的近似能够准确捕捉概念对损失景观的真实因果效应，从而验证了 CCBM 是一种可靠的模型行为解释工具。

CCBMs 可以消除数据影响。 为了验证数据级可控性（特别是机器遗忘）的有效性，我们采用了成员推断攻击（MIA）。在符合隐私要求的机器遗忘场景中，被移除的训练样本应与模型从未见过的样本（非成员）无法区分。

为了对此进行严格量化，我们使用**RMIA（移除成员推断攻击）分数**。RMIA 分数近似表示样本属于训练集的似然与其属于非成员分布的似然之比，定义如下：

$$LR_\theta(x,z) \approx \frac{\Pr(f_\theta(x)|\mathcal{N}(\mu_{x,\bar{z}}(x),\sigma^2_{x,\bar{z}}(x)))}{\Pr(f_\theta(x)|\mathcal{N}(\mu_{\bar{x},z}(x),\sigma^2_{\bar{x},z}(x)))} \times \frac{\Pr(f_\theta(z)|\mathcal{N}(\mu_{x,\bar{z}}(z),\sigma^2_{x,\bar{z}}(z)))}{\Pr(f_\theta(z)|\mathcal{N}(\mu_{\bar{x},z}(z),\sigma^2_{\bar{x},z}(z)))}, \tag{17}$$

其中 $f_\theta(x)$ 表示模型的 logits，而 $\mathcal{N}(\cdot)$ 表示针对成员/非成员假设拟合的高斯分布 [36]。RMIA 分数越高，表明样本属于训练集成员的概率越高。

我们计算了 200 名训练成员和 200 名非成员的 RMIA 分数。频率分布绘制于图 5a （编辑前）和图 5b （编辑后）。

- **编辑前（图 5a）：** 两种分布之间存在明显的分离。与非成员（绿色曲线，$\mu = 0.0495$）相比，成员（蓝色柱，红色拟合曲线）的 RMIA 分数显著更高（均值 $\mu = 0.0635$）。这种分离体现了模型中存储的“隐私泄露”或数据足迹。

- **编辑后（图 5b）：** 应用 CCBM 对成员样本进行机器遗忘后，被移除成员的分布明显向左偏移。新的 RMIA 分数均值降至 $\mu = 0.0521$，逐渐趋近非成员基线。从图中可以看出，红色曲线（被移除的成员）现在与绿色曲线（非成员）显著重叠。

这一统计变化证实，CCBM成功“净化”了模型。被移除的样本不再触发高置信度的成员身份信号，从而有效缓解隐私风险，并满足遗忘要求。

5.4 通过有害数据删除进行改进

为严格评估CCBM通过数据损坏恢复模型性能的能力，我们按照“噪声注入与移除”协议，在CUB数据集上开展了额外实验。具体而言，我们以三种不同粒度向训练集中合成引入了噪声：

- **概念层级：** 我们随机选取了 10% 的概念属性，并将其二元值翻转，应用于部分训练数据。

- **数据级别：**我们随机选取了 10% 的训练样本，并对其真实分类标签（y）进行了扰动。
- **概念标签层级：**我们通过在整个数据集中随机翻转 10 % 的单个概念标注（c_{ij}），引入了细粒度噪声。

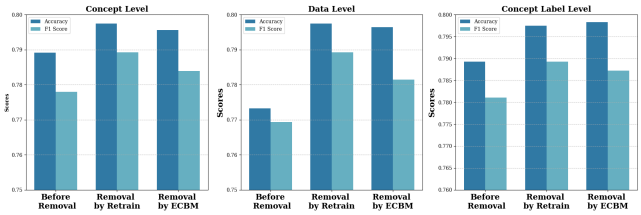


图 6：移除有害数据后的模型性能。

注入噪声后，我们在损坏数据集上训练了一个基线概念瓶颈模型（“移除前”）。随后，我们识别出噪声成分，并通过两种方法将其移除：(1) **从头开始重新训练**，使用清理后的数据集（作为真实值）；(2) **CCBM**，直接编辑已训练的含噪模型，使其遗忘有害组件。

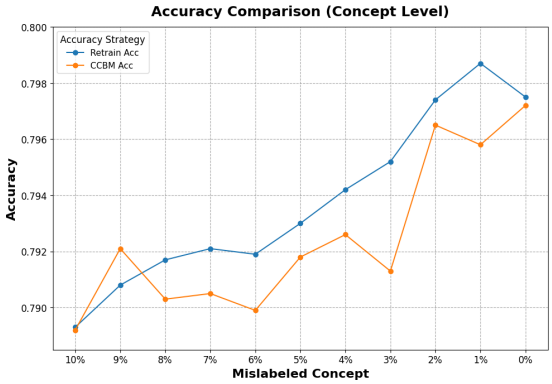
结果如图6所示。下面我们将详细分析这些发现：

性能显著恢复。从“移除前”的柱状图可以看出，注入噪声明显降低了模型效用。这种影响在数据级别最为显著，标签损坏使准确率降至约 0.773。然而，应用 CCBM 移除这些有害样本后，模型性能显著恢复，准确率回升至 ≈0.796，F1 分数则从 ≈0.769 提高到 ≈0.782。在概念层级和概念标签层级也观察到了类似的恢复趋势，证实 CCBM 能有效消除错误数据带来的负面影响。

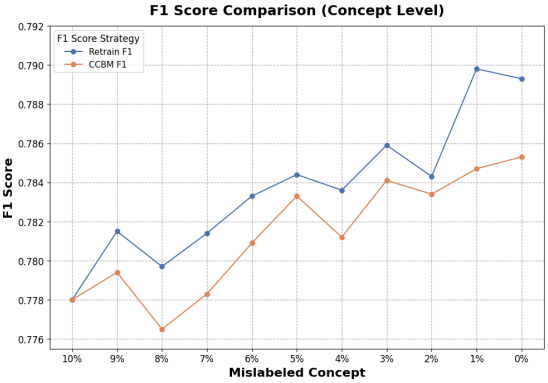
与重新训练的结果高度一致。可编辑模型的一项关键标准是其近似黄金标准重新训练结果的能力。图 6 显示，CCBM 能够以很高的保真度实现这一点。在全部三种设置中，“通过CCBM移除”与“通过重新训练移除”之间的性能差距都微乎其微。例如，在概念标签层级，F1 分数的差异小于 0.003。虽然与完整重新训练相比，性能略有下降（这归因于影响函数的近似误差），但 CCBM 保留了 99% 以上的恢复收益。这表明，CCBM 可作为重新训练的一种稳健且高效的替代方案，用于数据清理任务，无需从头训练所带来的计算开销，即可快速纠正数据集错误。

5.5 周期性编辑的性能

在现实世界的生命周期管理中，模型并非只编辑一次，而是会持续进行定期更新。在此类场景中，一个关键挑战是误差累积：由于影响函数提供的是参数变化的一阶近似，反复应用这些近似理论上可能会导致模型参数随着时间推移逐渐偏离最优模型参数。



(a) 编辑后的模型与重新训练模型的准确率对比。



(b) 编辑后的模型与重新训练。

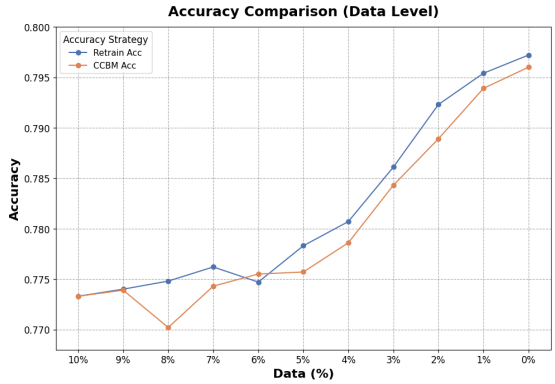
图 7：编辑后的模型与重新训练模型在概念层级上的准确率和 F1 分数差异。

为了严格评估 CCBM 在序贯编辑下的稳定性，我们设计了“多阶段恢复”实验。**设置：**我们首先按照上一节中的相同设置，在 CUB 数据集的三个层级（概念、数据和概念标签）中注入 10% 的噪声。**流程：**我们没有一次性移除所有噪声，而是执行了由 10 轮组成的序贯清理过程。在每一轮中，我们识别并移除/纠正 1% 的噪声数据（或概念）。关键在于，对于 CCBM方法，第 t 轮的更新直接应用于第 $t-1$ 轮产生的近似模型，从而测试该方法对累积近似误差的抵抗能力。

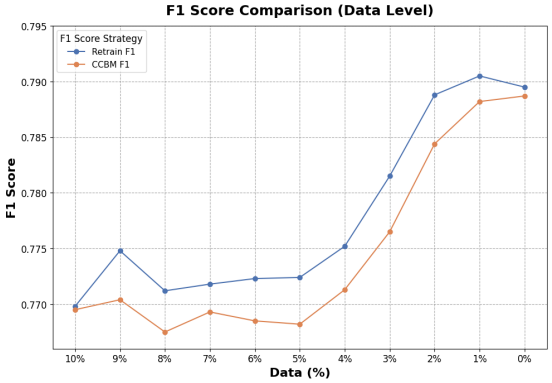
这 10 轮中准确率和 F1 分数的变化轨迹分别绘制在图 7（概念层级）、图 8（数据级别）和图9（概念标签层级）中。我们观察到以下主要趋势：

稳定的单调恢复。随着噪声从 10%（x 轴左侧）降至 0%（x 轴右侧），重新训练（蓝色）和 CCBM（橙色）的性能都呈现稳定的上升趋势。例如，在数据级别（图8）中，准确率从 ∼0.773 提升至∼0.796。这证实了 CCBM 能够有效地将有害组件的移除转化为每一步都能实现的切实性能提升。

对误差累积的鲁棒性。最重要的发现是 CCBM 轨迹的稳定性。尽管连续更新了 10 次，CCBM 性能曲线仍与重新训练曲线保持平行，没有显著

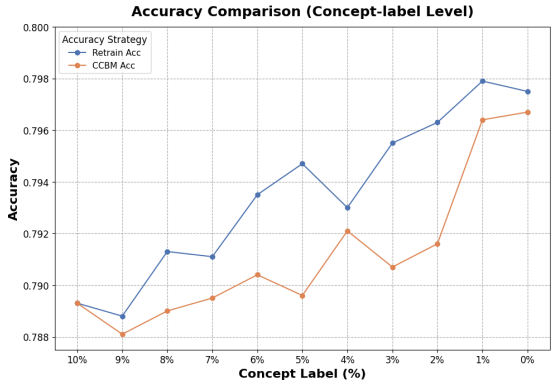


(a) 编辑后的模型与重新训练后的模型相比的准确率。

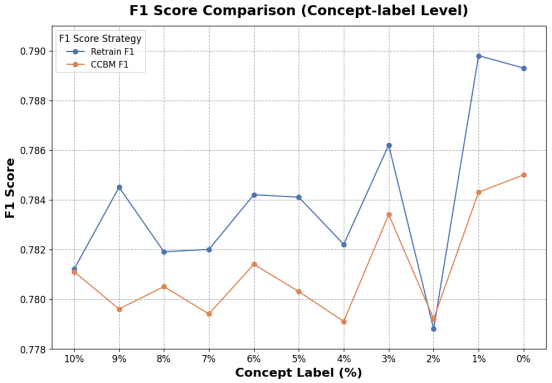


(b) 编辑后的模型与重新训练。

图8：在数据级别上，编辑后的模型与重新训练后的模型在准确率和 F1 分数方面的差异。



（a）编辑后的模型与重新训练模型相比的准确率。



（b）编辑后的模型与重新训练模型相比的 F1 分数。

图 9：编辑后的模型与重新训练模型在概念标签级上的准确率和 F1 分数差异。

差异。在数据级别（图8）上，CCBM 曲线紧贴重新训练曲线，最终准确率差距可以忽略不计（< 0.002）。在概念层级（图7）上，尽管概念移除问题的难度导致存在系统性偏移，但这一差距在整个 10 个阶段中始终保持不变。CCBM 模型不会崩溃，也不会表现出不稳定行为。

这一实证证据表明，CCBM 中基于海森矩阵的近似具有足够的鲁棒性，能够支持模型的长期定期维护，而无需频繁进行完整的重新训练重置。

6 结论

本文提出了**可控概念瓶颈模型（CCBMs）**，这是一个旨在解决传统可解释模型静态性这一关键局限的统一框架。鉴于现实世界的部署需要持续适应，我们围绕三种不同粒度，对模型部署后的可控性问题进行了形式化：概念标签级用于纠正标注错误，概念层级用于演化语义本体，以及数据级 用于管理训练数据生命周期。值得注意的是，我们的框架支持双向数据操作，既能严格执行对敏感或有害样本的机器遗忘，又能高效添加新数据以进行增量学习。

在方法论方面，我们基于影响函数推导出数学上严谨的闭式近似，从而能够精确更新参数，而无需付出从头重新训练的高昂代价。为确保适用于高维模型，我们进一步整合了特征值校正的 Kronecker 因子近似曲率（EK-FAC），以加速海森矩阵计算。

在多个基准测试上开展的大量实验表明，CCBMs 在效用与效率之间实现了更优的权衡。我们的方法所实现的模型更新在理论上与真实值重训练保持一致，同时将计算时间缩短了数个数量级。此外，我们通过成员推断攻击实证验证了该方法在隐私方面的优势，证实其能够有效消除数据痕迹。我们相信，CCBMs 将为更加可持续、自适应且可信赖的可解释人工智能系统铺平道路，使其能够与人类知识和动态数据环境共同演进。

7 影响声明

这项研究改进了对比学习中的数据归因方法。虽然我们承认评估社会影响的重要性，但我们的分析表明，除机器学习中的标准实践外，目前没有需要采取具体缓解措施的紧迫伦理风险。

参考文献

[1] M. A. Ahmad, C. Eckert, and A. Teredesai, 《医疗保健中的可解释机器学习》，载于《2018年ACM生物信息学、计算生物学与健康信息学国际会议论文集》，2018年，第559–560页。 [2] K.-H. Yu, A. L. Beam, and I. S. Kohane, 《医疗保健中的人工智能》，《Nature Biomedical Engineering》，第2卷，731，2018年10月，第719页。[3] L. Cao, 《金融领域的人工智能：挑战、技术与机遇》，《ACM计算调查（CSUR）》，第55卷，第3期，第1–38页，2022年。 [4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong 等, 《大语言模型综述》，arXiv预印本 arXiv:2303.18223，2023年。 [5] S. Yang, M. A. Ali, C.-L. Wang, L. Hu, and D. Wang, 《MORAL：面向大语言模型终身学习的 MoE 增强 LoRA》，arXiv预印本 arXiv:2402.11260，2024年。 [6] S. Yang, L. Hu, L. Yu, M. A. Ali, and D. Wang, 《通信时代的人机交互：自噬使大型模型达到局部最优》，arXiv预印本 arXiv:2402 11271，2024年。 [7] X. Xu, K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli, 《大语言模型会欺骗自己：一种基于提示的对抗攻击》，arXiv预印本 arXiv:2310.13345，2023年。 [8] S. Yang, J. Su, H. Jiang, M. Li, K. Cheng, M. A. Ali, L. Hu, and D. Wang, 《辩证对齐：化解 3h 的张力与大语言模型的安全威胁》，arXiv预印本 arXiv:2404.00486，2024年。 [9] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, 《多模态大语言模型综述》，arXiv预印本 arXiv:2306.13549，2023年。 [10] M. A. Ali, Z. Li, S. Yang, K. Cheng, Y. Cao, T. Huang, L. Hu, L. Yu, and D. Wang, 《Prompt-SAW：利用关系感知图进行文本提示压缩》，arXiv预印本 arXiv:2404.00489，2024年。 [11] K. Cheng, G. Lin, H. Fei, L. Yu, M. A. Ali, L. Hu, D. Wang 等, 《时序知识编辑下的多跳问答》，arXiv预印本 arXiv:2404.00492，2024年。 [12] A. Das and P. Rad, 《可解释人工智能（XAI）的机遇与挑战：综述》，arXiv预印本 arXiv:2006 11371，2020年。 [13] L. Hu, Y. Liu, N. Liu, M. Huai, L. Sun, and D. Wang, 《SEAT：稳定且可解释的注意力》，载于《AAAI人工智能会议论文集》，第37卷（第11期），2023年，第12 907–12 915页。[14] ——, 《提升视觉变换器的忠实性》，arXiv预印本 arXiv:2311.17983，2023年。 [15] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, 《概念瓶颈模型》，载于《国际机器学习会议》。PMLR，2020年，第5338–5348页。 [16] T. Oikarinen, S. Das, L. Nguyen, and L. Weng, 《无标签概念瓶颈模型》，载于《国际学习表征会议》，2023年。 [17] M. Yuksekgonul, M. Wang, and J. Zou, 《后处理概念瓶颈模型》，载于《第十一届国际学习表征会议》，2023年。 [18] S. Lai, L. Hu, J. Wang, L. Berti-Equille, and D. Wang, 《通过概念瓶颈模型实现忠实的视觉语言解释》，载于《第十二届国际学习表征会议》，2023年。 [19] I. Sheth and S. Ebrahimi Kahou, 《用于学习可泛化的基于概念的模型的辅助损失》，载于《神经信息处理系统进展》，A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine 编，第36卷，2023年，第26 966–26 990页。 [20] M. Espinosa Zarlenga, P. Barbiero, G. Ciravegna, G. Marra, F. Giannini, M. Diligenti, Z. Shams, F. Precioso, S. Melacci, A. Weller, P. Li´o, and M. Jamnik, 《概念嵌入模型：超越准确性与可解释性权衡》，载于《神经信息处理系统进展》，第35卷，2022年。 [21] Y. Rozenfeld, J. Beam, H. Maier, W. Haggerson, K. Boudreau, J. Carlson, and R. Medows, 《差异模型：与新冠病毒感染相关的风险因素》，《国际健康公平期刊》，第19卷，第1期，第126页，2020年。 [22] R. D. Cook, 《线性回归中有影响力观测值的检测》，《Technometrics》，第42卷，第1期，第65–68页，2000年。 [23] R. D. Cook and S. Weisberg, 《用于检测回归中有影响力案例的经验影响函数特征》，《Technometrics》，第22卷，第4期，第495–508页，1980年。 [24] K. Chauhan, R. Tiwari, J. Freyberg, P. Shenoy, and K. Dvijotham, 《交互式概念瓶颈模型》，载于《AAAI人工智能会议论文集》，第37卷（第5期），2023年，第5948–5955页。

[25] M. Vélez, 《解决概念瓶颈模型中的信息泄漏》，《神经信息处理系统进展》，第35卷，第23 386–23 397页，2022年。[26] E. Kim、D. Jung、S. Park、S. Kim和S. Yoon, 《概率概念瓶颈模型》，arXiv预印本 arXiv:2306.01574，2023年。[27] M. Keser、G. Schwalbe、A. Nowzad和A. Knoll, 《利用领域不变概念瓶颈模型对二维目标检测器进行可解释的模型无关合理性验证》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2023年，第3890–3899页。[28] Y. Sawada和K. Nakamura, 《包含额外无监督概念的概念瓶颈模型》，IEEE Access，第10卷，第41 758–41 765页，2022年。[29] I. Sheth和S. E. Kahou, 《用于学习泛化能力强的基于概念模型的辅助损失》，载于第三十七届神经信息处理系统大会，2023年。[30] M. Bartlett, 《近似置信区间》，Biometrika，第40卷，第1/2期，第12–19页，1953年。[31] M. Nevitt、D. Felson和G. Lester, 《骨关节炎倡议》，队列研究方案，第1卷，第2页，2006年。[32] C. Wah、S. Branson、P. Welinder、P. Perona和S. Belongie, 《Caltech-UCSD Birds-200-2011数据集》，加州理工学院，2011年。[33] Z. Liu、P. Luo、X. Wang和X. Tang, 《野外人脸属性的深度学习》，载于IEEE国际计算机视觉会议论文集，2015年，第3730–3738页。[34] J. Kawahara、S. Daneshvar、G. Argenziano和G. Hamarneh, 《使用多任务多模态神经网络进行七点检查表评估和皮肤病变分类》，IEEE生物医学与健康信息学杂志，第23卷，第2期，第538–546页，2019年3月。[35] R. Daneshjou、M. Yuksekgonul、Z. R. Cai、R. Novoa和J. Zou, 《Skincon：由领域专家密集标注的皮肤病数据集，用于细粒度模型调试与分析》，载于《第36届神经信息处理系统国际会议论文集》，NIPS '22系列。美国纽约州雷德胡克：Curran Associates Inc.，2022年。[36] S. Zarifzadeh、P. Liu和R. Shokri, 《低成本、高效能的成员推理攻击》，载于第四十一届国际机器学习会议，2024年。[37] T. George、C. Laurent、X. Bouthillier、N. Ballas和P. Vincent, 《在克罗内克因子化特征基中快速近似自然梯度下降》，《神经信息处理系统进展》，第31卷，2018年。[38] P. W. Koh和P. Liang, 《通过影响函数理解黑箱预测》，载于国际机器学习会议。PMLR，2017年，第1885–1894页。[39] A. Warnecke、L. Pirch、C. Wressnegger和K. Rieck, 《特征与标签的机器遗忘》，网络与分布式系统安全（NDSS）研讨会，2023年。[40] J. Liu、J. Lou、Z. Qin和K. Ren, 《具有泛化速率和删除容量的认证极小极大遗忘》，《神经信息处理系统进展》，第36卷，2024年。[41] A. Golatkar、A. Achille和S. Soatto, 《纯净网络的永恒阳光：深度网络中的选择性遗忘》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第9304–9312页。[42] A. Golatkar、A. Achille、A. Ravichandran、M. Polito和S. Soatto, 《深度网络中的混合隐私遗忘》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第792–801页。[43] N. Agarwal、B. Bullins和E. Hazan, 《用于机器学习的线性时间二阶随机优化》，《机器学习研究杂志》，第18卷，第116期，第1–40页，2017年。[44] Y. Kwon、E. Wu、K. Wu和J. Zou, 《Datainf：高效估算LoRA微调的大语言模型和扩散模型中的数据影响》，载于第十二届国际学习表征会议，2023年。[45] H. Guo、N. Rajani、P. Hase、M. Bansal和C. Xiong, 《FastIF：用于高效模型解释与调试的可扩展影响函数》，载于《2021年自然语言处理实证方法会议论文集》，2021年，第10 333–10 350页。[46] X. Han、B. C. Wallace和Y. Tsvetkov, 《解释黑箱预测并通过影响函数揭示数据伪影》，arXiv预印本 arXiv:2005.06676，2020年。[47] S. Basu、P. Pope和S. Feizi, 《深度学习中的影响函数并不稳定》，载于国际学习表征会议（ICLR），2021年。

[48] H. Wang、B. Ustun和F. Calmon，《无需重新训练进行修复：利用反事实分布避免差别影响》，载于国际机器学习会议。PMLR，2019年，第6618–6627页。[49] M.-E. Brunet、C. Alkalay-Houlihan、A. Anderson和R. Zemel，《理解词嵌入中偏置的起源》，载于国际机器学习会议。PMLR，2019年，第803–811页。[50] H. Chen、S. Si、Y. Li、C. Chelba、S. Kumar、D. Boning和C.-J. Hsieh，《多阶段影响函数》，神经信息处理系统进展，第33卷，第12 732–12 742页，2020年。[51] L. Bourtoule、V. Chandrasekaran、C. A. Choquette-Choo、H. Jia、A. Travers、B. Zhang、D. Lie和N. Papernot，《机器遗忘》，载于*2021*年*IEEE*安全与隐私研讨会（*SP*）。IEEE，2021年，第141–159页。[52] J. Brophy和D. Lowd，《随机森林的机器遗忘》，载于国际机器学习会议。PMLR，2021年，第1092–1104页。[53] Y. Cao和J. Yang，《迈向利用机器遗忘使系统学会遗忘》，载于*2015*年*IEEE*安全与隐私研讨会。IEEE，2015年，第463–480页。[54] C. Chen、F. Sun、M. Zhang和B. Ding，《推荐遗忘》，载于*2022*年*ACM Web*会议论文集，2022年，第2768–2777页。[55] M. Chen、Z. Zhang、T. Wang、M. Backes、M. Humbert和Y. Zhang，《图遗忘》，载于*2022*年*ACM SIGSAC*计算机与通信安全会议论文集，2022年，第499–513页。[56] A. Sekhari、J. Acharya、G. Kamath和A. T. Suresh，《记住你想遗忘的内容：机器遗忘算法》，神经信息处理系统进展，第34卷，第18 075–18 086页，2021年。[57] S. B. R. Chowdhury、K. Choromanski、A. Sehanobish、A. Dubey和S. Chaturvedi，《利用参数高效微调实现可扩展的精确机器遗忘》，*arXiv*预印本 *arXiv:2406 16257*，2024年。[58] C. Guo、T. Goldstein、A. Hannun和L. Van Der Maaten，《机器学习模型的认证数据移除》，*arXiv*预印本 *arXiv:1911.03030*，2019年。[59] Z. Izzo、M. A. Smart、K. Chaudhuri和J. Zou，《从机器学习模型中近似删除数据》，载于国际人工智能与统计会议。PMLR，2021年，第2008–2016页。[60] R. Mehta、S. Pal、V. Singh和S. N. Ravi，《通过随机条件独立海森矩阵实现深度遗忘》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第10 422–10 431页。[61] A. Golatkar、A. Achille和S. Soatto，《跳出框架遗忘：清除深度网络中可通过输入—输出观测获取的信息》，载于《计算机视觉—*ECCV 2020*：第*16*届欧洲会议，英国格拉斯哥，*2020*年*8*月*23*日至*28*日，论文集，第*29*部分 *16*》。施普林格，2020年，第383–398页。[62] Y. Wu、E. Dobriban和S. Davidson，《DeltaGrad：机器学习模型的快速重新训练》，载于国际机器学习会议。PMLR，2020年，第10 355–10 366页。[63] A. Thudi、G. Deza、V. Chandrasekaran和N. Papernot，《展开SGD：理解影响机器遗忘的因素》，载于*2022*年第七届*IEEE*欧洲安全与隐私研讨会（*EuroS&P*）。IEEE，2022年，第303–319页。



HongbinLin 于2025年获得香港科技大学（中国香港）的人工智能哲学硕士学位。目前，他是穆罕默德·本·扎耶德人工智能大学（MBZUAI）机器学习系的研究助理。他的研究兴趣包括多模态大语言模型和深度学习理论。



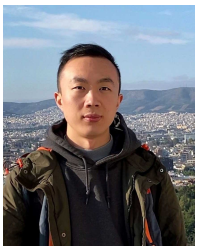
ChenyangRen于2025年获得上海交通大学（中国上海）的应用数学理学硕士学位。目前，她正在香港科技大学中国香港特别行政区校区数学系攻读博士学位。她的研究兴趣包括材料科学人工智能、机器学习和科学计算。



JuanguiXu于2024年获得中国广州大学网络空间安全工学学士学位。目前正在德国萨尔大学语言科学与技术系攻读理学硕士学位。他的研究兴趣包括计算机视觉和大语言模型。



Zhengyu Hu是香港科技大学（广州）人工智能专业的哲学硕士研究生，师从Hui Xiong教授。他的研究领域涵盖机器学习和数据挖掘，近期重点关注大语言模型、偏好评估、多智能体系统和图学习。他已在NeurIPS、ICML、KDD、EMNLP、ICCV和WWW等顶级会议上发表十余篇论文。他还担任KDD、NeurIPS等重要会议的审稿人，ICLR、ICML、CVPR、ACL、AAAI、AISTATS、WWW、NAACL、COLING、ECCV 和 EMNLP，以及 DMLR 和 IEEE TNNLS 等期刊。



Cheng-LongWang是 KAUST 计算机科学专业的博士生，现处于最后一年，由 Di Wang 博士指导，并与 Yinzhi Cao 博士密切合作。他的研究重点是机器遗忘、数据/大语言模型记忆，以及相关的隐私、版权和安全挑战。



Yao Shu于2017年获得中国华中科技大学工学学士学位，并于2022年获得新加坡新加坡国立大学博士学位。他目前是中国香港科技大学（广州）人工智能学域的长聘轨助理教授。其研究兴趣包括深度学习理论和优化算法。



Hui Xiong（IEEE会士）获得明尼苏达大学计算机科学博士学位。他是中国香港科技大学（广州）讲席教授、副校长（知识转移）及人工智能学域教授。他还任职于中国香港香港科技大学计算机科学与工程系。其研究兴趣涵盖人工智能、数据挖掘和移动计算。

他曾在众多会议的组织委员会和程序委员会任职，包括担任第18届ACM SIGKDD 知识发现与数据挖掘国际会议（KDD）工业与政府专题的程序联合主席、2013年 IEEE 国际数据挖掘会议（ICDM）的程序联合主席、2015年 IEEE 国际数据挖掘会议（ICDM）的总联合主席，以及2018年 ACM SIGKDD 知识发现与数据挖掘国际会议研究专题的程序联合主席。他曾获得多项奖项，包括2021年 AAAI 最佳论文奖和2011年 IEEE ICDM 最佳研究论文奖。由于在数据挖掘和移动计算领域作出的重要贡献，他于2020年当选为美国科学促进会会士。



JingfengZhang （IEEE会员）于2016年获得中国济南山东大学泰山学堂计算机科学学士学位，并于2020年获得新加坡国立大学计算机科学博士学位。2021年至2022年，他在日本中央区理化学研究所先进智能项目中心（AIP）（理化学研究所 AIP）担任博士后研究员，并于2023年成为研究科学家。他目前是讲师，也是博士生认证导师，任职于 The

奥克兰大学，并兼任理化学研究所 AIP 的访问科学家。张博士于2022年获得了JST奖、日本学术振兴会科学研究费助成金（KAKENHI）青年科学家项目资助，以及理化学研究所大武奖。



Di Wang 自2021年春季起，担任阿卜杜拉国王科技大学（KAUST）计算机、电气与数学科学及工程学部（CEMSE）的计算机科学助理教授，并兼任统计学领域的教职人员。他还是可证明负责任人工智能与数据分析（PRADA）实验室的首席研究员，也是生成式人工智能卓越中心的成员。在此之前，他在纽约州立大学（SUNY）布法罗分校攻读计算机科学博士学位，

并于2020年在Jinhui Xu博士的指导下获得学位。在攻读博士学位之前，我于2015年在西安大略大学获得数学硕士学位；他于2014年在山东大学获得数学与应用数学学士学位。



Lijie Hu是穆罕默德·本·扎耶德人工智能大学（MBZUAI）机器学习系的助理教授。她于2025年在阿卜杜拉国王科技大学（KAUST）获得计算机科学博士学位，期间有幸师从王迪教授。王迪教授领导着PRADA实验室（可证明负责任人工智能与数据分析实验室）。在此之前，她获得了中国人民大学的数学硕士学位。

8 影响函数

考虑一个神经网络 $\hat{\theta} = \arg\min_\theta \sum_{i=1}^n \ell(z_i, \theta)$，其损失函数为 L，数据集为 $D = \{z_i\}_{i=1}^n$。也就是说，θ 最小化经验风险

$$R(\theta) = \sum_{i=1}^n L(z_i, \theta)$$

假设 R 在 θ中是强凸的。那么 θ 有唯一确定的定义。如果我们从训练数据集中移除一个点 z_m，则参数变为 $\theta_{-z_m} = \arg\min_\theta \sum_{i\neq m} L(z_i, \theta)$。将 z_m 的权重增加足够小的 ϵ 后，修订后的风险 $R(\theta)' = \frac{1}{n}\sum_{i=1}^n L(z_i;\theta) + \epsilon L(z_m;\theta)$仍然是强凸的。那么响应函数 $\theta_{\epsilon,-z_m} = R(\theta)$也有唯一确定的定义。参数变化记为 $\Delta_\epsilon = \theta_{\epsilon,-z_m} - \theta$。由于 $\theta_{\epsilon,-z_m}$ 是 $R(\theta)$的极小化器，我们得到一阶优化条件如下

$$\nabla_{\hat{\theta}_{\epsilon,-z_m}} R(\theta) + \epsilon \cdot \nabla_{\hat{\theta}_{\epsilon,-z_m}} L(z_m, \hat{\theta}_{\epsilon,-z_m}) = 0$$

由于 $\theta_{\epsilon,-z_m} \to \theta as \epsilon \to 0$，我们对右侧进行泰勒展开：

$$\left[\nabla R(\hat{\theta}) + \epsilon \nabla L(z_m, \hat{\theta})\right] + \left[\nabla^2 R(\hat{\theta}) + \epsilon \nabla^2 L(z_m, \hat{\theta})\right] \Delta_\epsilon \approx 0$$

注意到 $\epsilon\nabla^2 L(z_m,\hat{\theta})\Delta_\epsilon$ 是 $o(\|\Delta_\epsilon\|)$ 项，其量级小于其他部分，因此在后续分析中将其舍去。于是，泰勒展开方程变为

$$\left[\nabla R(\hat{\theta}) + \epsilon \nabla L(z_m, \hat{\theta})\right] + \nabla^2 R(\hat{\theta}) \cdot \Delta_\epsilon \approx 0$$

求解 Δ_ϵ，得到：

$$\Delta_\epsilon = -\left[\nabla^2 R(\hat{\theta}) + \epsilon \nabla^2 L(z, \hat{\theta})\right]^{-1} \left[\nabla R(\hat{\theta}) + \epsilon \nabla L(z, \hat{\theta})\right].$$

记住，θ 使 R 最小化，则 $\nabla R(\theta) = 0$。舍去 $o(\epsilon)$项后，得到

$$\Delta_\epsilon = -\epsilon \nabla^2 R(\hat{\theta})^{-1} \nabla L(z, \hat{\theta}).$$

$$\left.\frac{d\hat{\theta}_{\epsilon,-z_m}}{d\epsilon}\right|_{\epsilon=0} = \left.\frac{d\Delta_\epsilon}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla L(z, \hat{\theta}) \equiv \mathcal{I}_{up,params}(z).$$

此外，我们还可以直接通过 $\theta_{-z_m} \approx \theta + \mathcal{I}_{up,params}(z)$ 得到 θ_{-z_m} 的近似值。

9 影响函数的加速

9.0.0.1 EK-FAC：EK-FAC 方法依赖于对费 Fisher 信息矩阵的两种近似，这在我们的设定中等同于 $G_{\hat{\theta}}$，从而使得计算该矩阵的逆矩阵成为可行。

首先，假设不同层中权重的导数彼此不相关，这意味着 $G_{\hat{\theta}}$ 具有块对角结构。假设 g_θ 可以表示为 $g_\theta(x) = g_{\theta_L} \circ \cdots \circ g_{\theta_l} \circ \cdots \circ g_{\theta_1}(x)$，其中 $l \in [L]$。我们将偏置并入权重，并将第 l 层的参数向量化为向量 $\theta_l \in \mathbb{R}^{d_l}$，$d_l \in \mathbb{N}$ 是第 l 层参数的数量。然后，$G_{\hat{\theta}}$ 可以替换为 $\left(G_1(\hat{\theta}) \cdots, G_L(\hat{\theta})\right)$，其中 $G_l(\hat{\theta}) \triangleq n^{-1}\sum_{i=1}^n \nabla_{\hat{\theta}_l} \ell_i \nabla_{\theta_l} \ell_i^{\mathrm{T}}$。

将 h_l、o_l 分别记为第 l 层的输出和预激活输出。则 $G_l(\theta)$ 可以近似为

$$G_l(\theta) \approx \hat{G}_l(\theta) \triangleq \frac{1}{n}\sum_{i=1}^n h_{l-1}(x_i)\, h_{l-1}(x_i)^T \otimes \frac{1}{n}\sum_{i=1}^n \nabla_{o_l} \ell_i \nabla_{o_l} \ell_i^T \triangleq \Omega_{l-1} \otimes \Gamma_l.$$

此外，为了加速转置运算并引入阻尼项，对矩阵 Ω_{l-1} 和 Γ_l 进行特征值分解，得到相应的分解结果 $Q_\Omega \Lambda_\Omega Q_\Omega^\top$ 和 $Q_\Gamma \Lambda_\Gamma Q_\Gamma^\top$。然后，$H_l(\theta)$ 的逆矩阵可由下式得到

$$\begin{aligned}\hat{H}_l(\theta)^{-1} &\approx \left(\hat{G}_l(\hat{g}) + \lambda_l I_{d_l}\right)^{-1} \\ &= \left(Q_{\Omega_{l-1}} \otimes Q_{\Gamma_l}\right)\left(\Lambda_{\Omega_{l-1}} \otimes \Lambda_{\Gamma_l} + \lambda_l I_{d_l}\right)^{-1} \\ &\quad \times \left(Q_{\Omega_{l-1}} \otimes Q_{\Gamma_l}\right)^{\mathrm{T}}.\end{aligned}$$

此外，[37] 提出了一种新方法，用于修正方程 18 中的误差，该方法将 i-th 对角元素 $\Lambda_{\Omega_{l-1}} \otimes \Lambda_{\Gamma_l}$ 设为 $\Lambda_{ii}^* = n^{-1}\sum_{j=1}^n \left(\left(Q_{\Omega_{l-1}} \otimes Q_{\Gamma_l}\right)\nabla_{\theta_l}\ell_j\right)_i^2$.

9.1 概念瓶颈模型的 EK-FAC

在我们的概念瓶颈模型中，标签预测器是一个单独的线性层，计算海森矩阵的成本也可以接受。然而，概念预测器基于 ResNet-18，参数数量很多。因此，我们对 g 使用 EK-FAC。

$$\hat{g} = \arg\min_g \sum_{j=1}^k L_{C_j} = \arg\min_g \sum_{j=1}^k \sum_{i=1}^n L_C(g^j(x_i), c_i^j),$$

我们将 $H_{\hat{g}} = \nabla_{\hat{g}}^2 \sum_{i,j} L_{C_j}(g(x_i)\ c_i)$ 定义为损失函数相对于参数的海森矩阵。

为此，考虑 g 的第 l 层，该层以激活层 $\{a_{j,t}\}$ 作为输入，其中 $j \in \{1, 2, \ldots, J\}$ 表示输入映射的索引，$t \in \mathcal{T}$ 表示空间位置的索引，通常为空间二维网格。该层由一组权重 $W = (w_{i,j,\delta})$ 和偏置 $b = (b_i)$ 参数化，其中 $i \in \{1, \ldots, I\}$ 表示输出映射的索引，$\delta \in \Delta$ 表示空间偏移的索引（相对于滤波器中心）。

卷积层计算一组预激活值，如下所示

$$[S_l]_{i,t} = s_{i,t} = \sum_{\delta\in\Delta} w_{i,j,\delta} a_{j,t+\delta} + b_i.$$

将关于 $s_{i,\ t}$ 的损失导数记为

$$\mathcal{D}s_{i,t} = \frac{\partial \sum L_{C_j}}{\partial s_{i,t}},$$

它可以在反向传播过程中计算得到。

激活值实际存储为维度为 $|\mathcal{T}| \times J$ 的 A_{l-1}。类似地，权重存储为一个 W_l $I \times |\Delta| J$ 数组。尽管理论上高度并行，直接实现卷积仍会受到较差内存访问模式的影响。相反，高效实现通常会利用一种称为扩展算子 $[\![\cdot]\!]$ 的技术。例如，$[\![A_{l-1}]\!]$ 是一个 $|\mathcal{T}| \times J|\Delta|$ 矩阵，定义如下：

$$[\![A_{l-1}]\!]_{t,j|\Delta|+\delta} = [A_{l-1}]_{(t+\delta),j} = a_{j,t+\delta},$$

为了将偏置并入权重，我们需要添加一个齐次坐标（即一列全为1的列）到

扩展后的激活值 $[\![A_{l-1}]\!]$ 中添加一个齐次坐标（即一列全为1的值），并将其记为 $[\![A_{l-1}]\!]_{\mathrm{H}}$。将偏置向量与权重矩阵拼接后，我们得到 $\theta_l = (b_l, W_l)$。

那么， $H_{\hat{g}}$ 的近似为：

$$\begin{aligned} G^{(l)}(\hat{g}) &= \mathbb{E}\left[\mathcal{D}w_{i,j,\delta}\mathcal{D}w_{i',j',\delta'}\right] \\ &= \mathbb{E}\left[\left(\sum_{t\in\mathcal{T}} a_{j,t+\delta}\mathcal{D}s_{i,t}\right)\left(\sum_{t'\in\mathcal{T}} a_{j',t'+\delta'}\mathcal{D}s_{i',t'}\right)\right] \\ &\approx \mathbb{E}\left[[\![A_{l-1}]\!]_{\mathrm{H}}^{\top}[\![A_{l-1}]\!]_{\mathrm{H}}\right] \otimes \frac{1}{|\mathcal{T}|}\mathbb{E}\left[\mathcal{D}S_l^{\top}\mathcal{D}S_l\right] \\ &\triangleq \Omega_{l-1}\otimes\Gamma_l. \end{aligned}$$

使用训练集的均值估计期望，

$$\begin{aligned} G^{(l)}(\hat{g}) &\approx \frac{1}{n}\sum_{i=1}^{n}\left([\![A_{l-1}^i]\!]_{\mathrm{H}}^{\top}[\![A_{l-1}^i]\!]_{\mathrm{H}}\right)\otimes\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{|\mathcal{T}|}\mathcal{D}S_l^{i\top}\mathcal{D}S_l^i\right) \\ &\triangleq \hat{\Omega}_{l-1}\otimes\hat{\Gamma}_l. \end{aligned}$$

此外，若因子 $\hat{\Omega}_{l-1}$ 和 $\hat{\Gamma}_l$ 的特征分解分别为 $Q_\Omega\Lambda_\Omega Q_\Omega^{\top}$ 和 $Q_\Gamma\Lambda_\Gamma Q_\Gamma^{\top}$，则 $\hat{\Omega}_{l-1}\otimes\hat{\Gamma}_l$ 的特征分解可写为：

$$\begin{aligned} \hat{\Omega}_{l-1}\otimes\hat{\Gamma}_l &= Q_\Omega\Lambda_\Omega Q_\Omega^{\top}\otimes Q_\Gamma\Lambda_\Gamma Q_\Gamma^{\top} \\ &= \left(Q_\Omega\otimes Q_\Gamma\right)\left(\Lambda_\Omega\otimes\Lambda_\Gamma\right)\left(Q_\Omega\otimes Q_\Gamma\right)^{\top}. \end{aligned}$$

由于后续需要进行求逆运算，且当前对 $G^{(l)}(\hat{g})$ 的近似为半正定，因此我们实际使用其阻尼版本：

$$\begin{aligned} \hat{G}^l(\hat{g})^{-1} &= \left(G_l(\hat{g}) + \lambda_l I_{d_l}\right)^{-1} \\ &= \left(Q_{\Omega_{l-1}}\otimes Q_{\Gamma_l}\right)\cdot\left(\Lambda_{\Omega_{l-1}}\otimes\Lambda_{\Gamma_l} + \lambda_l I_{d_l}\right)^{-1} \\ &\quad\cdot\left(Q_{\Omega_{l-1}}\otimes Q_{\Gamma_l}\right)^{\mathsf{T}}. \end{aligned} \tag{18}$$

此外， [37] 提出了一种新方法，用于修正方程18 该方法将 i-th diagonal elementof$\Lambda_{\Omega_{l-1}}\otimes\Lambda_{\Gamma_l}$ 设为

$$\Lambda_{ii}^{*} = n^{-1}\sum_{j=1}^{n}\left(\left(Q_{\Omega_{l-1}}\otimes Q_{\Gamma_l}\right)\nabla_{\theta_l}\ell_j\right)_i^2.$$

10 算法

算法 1概念-标签级CCBM

1: 输入： 数据集 $\mathcal{D} = \{(x_i,\ y_i,\ c_i)\}_{i=1}^{n}$、原始概念ˆ预测器 f以及标签预测器ˆg，一组错误数据D_e 及其关联索引集 S_e。2: 对于 S_e 中的索引$(w,\ r)$，将 w 个数据$(x_w,\ y_w,\ c_w)$的 c_w^r 更正为正确标签$c_w^r{}'$。3: 计算损失函数关于ˆg的海森矩阵：

$$H_{\hat{g}} = \nabla_{\hat{g}}^2 \sum_{i,j} L_{C_j}(\hat{g}^j(x_i), c_i^j).$$

4: 更新概念预测器 $\tilde{g}$:

$$\tilde{g} = \hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{(w,r)\in S_e} \Big(\nabla_{\hat{g}} L_{C_r}\left(\hat{g}^r(x_w), c_w^r{}'\right) - \nabla_{\hat{g}} L_{C_r}\left(\hat{g}^r(x_w), c_w^r\right)\Big).$$

5: 计算损失函数关于ˆf的海森矩阵：

$$H_{\hat{f}} = \nabla_{\hat{f}}^2 \sum_{i=1}^{n} L_{Y_i}(\hat{f}, \hat{g}).$$

˜6: 更新标签预测器 f:

$$\tilde{f} = \hat{f} + H_{\hat{f}}^{-1} \cdot \nabla_f \sum_{i=1}^{n} L_Y\left(\hat{f}\left(\hat{g}(x_i)\right), y_i\right) - H_{\hat{f}}^{-1} \cdot \nabla_f \sum_{l=1}^{n} \left(L_Y\left(\hat{f}\left(\tilde{g}(x_l)\right), y_l\right)\right).$$

˜7: **返回：** f，$\tilde{g}$。

算法 2概念层级 CCBM

1: 输入： 数据集 $\mathcal{D} = \{(x_i,\ y_i,\ c_i)\}_{i=1}^{n}$，原始概念ˆ预测器 f，标签预测器 ˆg 以及待移除概念索引集 M。2: 对于 $r \in M$，为所有数据 $z \in \mathcal{D}$设置 $p_r = 0$。3: 计算损失函数关于ˆg的海森矩阵：

$$H_{\hat{g}} = \nabla_{\hat{g}}^2 \sum_{j\notin M} \sum_{i=1}^{n} L_{C_j}(\hat{g}^j(x_i), c_i^j).$$

4: 更新概念预测器 $\tilde{g}^*$:

$$\tilde{g}^* = \hat{g} - H_{\hat{g}}^{-1} \cdot \nabla_{\hat{g}} \sum_{j\notin M} \sum_{i=1}^{n} L_{C_j}(\hat{g}^j(x_i), c_i^j).$$

5: 计算损失函数关于ˆ f的海森矩阵：

$$H_{\hat{f}} = \nabla_{\hat{f}}^2 \sum_{i=1}^{n} L_Y(\hat{f}(\hat{g}(x_i), y_i).$$

˜6: 更新标签预测器 f:

$$\tilde{f} = \hat{f} - H_{\hat{f}}^{-1} \cdot \nabla_{\hat{f}} \sum_{l=1}^{n} L_Y\left(\hat{f}\left(\tilde{g}^*(x_l)\right), y_l\right).$$

7: 通过移除 $r \in M$ 的 $\tilde{g}^*$ 最后一层中的矩阵的第 r 行，将 $\tilde{g}^*$ 映射 $\tilde{g}$。

˜8: **返回：** f，$\tilde{g}$。

算法3数据级 CCBM（移除）

1: 输入： 数据集 $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^{N}$，原始概念ˆ预测器 f、标签预测器ˆg，以及要移除的数据索引集 G。2: 对于 $r \in G$，从 $\mathcal{D}$ 中移除第 r 个数据(x_r, y_r, c_r)，并将新数据集定义为 $\mathcal{S}$。3: 计算损失函数关于ˆg的海森矩阵：

$$H_{\hat{g}} = \nabla_{\hat{g}}^2 \sum_{i,j} L_{C_j}(\hat{g}^j(x_i), c_i^j).$$

4: 更新概念预测器 $\tilde{g}$:

$$\tilde{g} = \hat{g} + H_{\hat{g}}^{-1} \cdot \sum_{r\in G} \nabla_g L_C(\hat{g}(x_r), c_r)$$

˜5: 更新标签预测器 f。计算损失函数关于 f 的海森矩阵ˆ:

$$H_{\hat{f}} = \nabla_{\hat{f}}^2 \sum_{i=1}^{n} L_Y(\hat{f}(\hat{g}(x_i), y_i).$$

6: 计算 A，如下所示:

$$A = H_{\hat{f}}^{-1} \cdot \sum_{i\in[n]-G} \nabla_{\hat{f}} L_Y\left(\hat{f}(\hat{g}(x_i)), y_i\right)$$

˜7: 将 f 设为

$$\bar{f} = \hat{f} + A$$

8：计算关于 ¯ 的损失函数关于 f 的海森矩阵：

$$H_{\bar{f}} = \nabla_{\bar{f}}^2 \sum_{i\in[n]-G} L_Y(\bar{f}(\hat{g}(x_i)), y_i).$$

9：计算 B 为

$$B = -H_{\bar{f}}^{-1} \cdot \sum_{i\in[n]-G} \nabla_{\hat{f}} \Big(L_Y\left(\bar{f}(\tilde{g}(x_i)), y_i\right) - L_Y\left(\bar{f}(\hat{g}(x_i)), y_i\right)\Big).$$

˜ ˜ˆ10：更新标签预测器 f为：$f = f + A + B$.

11：**返回：** $\tilde{f}$，$\tilde{g}$。

算法 4数据级 CCBM（添加）

1: 输入： 数据集 $\mathcal{D} = \{(x_i,\ y_i,\ c_i)\}_{i=1}^{N}$、原始概念ˆ预测器 f、标签预测器ˆg，以及要添加的数据集 $\tilde{\mathcal{D}} = \{(\tilde{x}_i,\ \tilde{y}_i, \tilde{c}_i)\}_{i=1}^{M}$。2: 计算损失函数相对于ˆ$g$的海森矩阵：

$$H_{\hat{g}} = \nabla_{\hat{g}}^2 \sum_{i,j} L_{C_j}(\hat{g}^j(x_i), c_i^j).$$

3：更新概念预测器 $\tilde{g}$:

$$\tilde{g} = \hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{i\in[M]} \nabla_g L_C(\hat{g}(\tilde{x}_i), \tilde{c}_i)$$

4：更新标签预测器 f。计算损失函数关于 f 的海森矩阵 $\hat{}$：

$$H_{\hat{f}} = \nabla^2_{\hat{f}} \sum_{i=1}^{n} L_Y(\hat{f}(\hat{g}(x_i), y_i).$$

5：计算 A，如下：

$$A = -H_{\hat{f}}^{-1} \cdot \left(\sum_{i\in[N]} \nabla_{\hat{f}} L_Y \left(\hat{f}(\hat{g}(x_i)), y_i \right) + \sum_{i\in[M]} \nabla_{\hat{f}} L_Y \left(\hat{f}(\hat{g}(\tilde{x}_i)), \tilde{y}_i \right) \right)$$

6：将 f 作为

$$\bar{f} = \hat{f} + A$$

7：计算关于 f的损失函数的海森矩阵：

$$H_{\bar{f}} = \nabla^2_{\bar{f}} \left(\sum_{i\in[N]} L_Y \left(\bar{f}(\hat{g}(x_i)), y_i \right) + \sum_{i\in[M]} L_Y \left(\bar{f}(\hat{g}(\tilde{x}_i)), \tilde{y}_i \right) \right).$$

8：计算 B，作为

$$B = -H_{\bar{f}}^{-1} \cdot \sum_{i\in[n]\setminus G} \nabla_{\hat{f}} \Big(L_Y(\bar{f}(\tilde{g}(x_i)), y_i) - L_Y(\bar{f}(\hat{g}(x_i)), y_i) \Big).$$

9：将标签预测器 f更新为：$f = f + A + B$.

10: **返回：** $\tilde{f}, \tilde{g}$。

算法5 针对概念预测器 g 的 EK-FAC

1: **输入：** 数据集 $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^N$，原始概念预测器 $\hat{g}$。

2: **对于** $\hat{g}$ 的第 l 个卷积层：**执行**
3: 定义该层的输入激活值 $\{a_{j,t}\}$、权重 $W = (w_{i,j,\delta})$和偏置 $b = (b_i)$;
4：获取扩展后的激活值 $[\![A_{l-1}]\!]$，如下：

$$[\![A_{l-1}]\!]_{t,j|\Delta|+\delta} = [A_{l-1}]_{(t+\delta),j} = a_{j,t+\delta},$$

5：计算预激活值：

$$[S_l]_{i,t} = s_{i,t} = \sum_{\delta\in\Delta} w_{i,j,\delta} a_{j,t+\delta} + b_i.$$

6：在反向传播过程中，获取 $\mathcal{D}s_{i,\ t}$，如下：

$$\mathcal{D}s_{i,t} = \frac{\partial \sum_{j=1}^{k} \sum_{i=1}^{n} L_{C_j}}{\partial s_{i,t}}$$

7：计算$\hat{\Omega}_{l-1}$ 和 $\hat{\Gamma}_l$:

$$\hat{\Omega}_{l-1} = \frac{1}{n} \sum_{i=1}^{n} \left([\![A^i_{l-1}]\!]^\top_{\mathrm{H}} [\![A^i_{l-1}]\!]_{\mathrm{H}} \right)$$
$$\hat{\Gamma}_l = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{|\mathcal{T}|} \mathcal{D}S^{i\top}_l \mathcal{D}S^i_l \right)$$

8：对$\hat{\Omega}_{l-1}$ 和 $\hat{\Gamma}_l$进行特征值分解，得到 Q_Ω，Λ_Ω，Q_Γ，Λ_Γ，满足

$$\hat{\Omega}_{l-1} = Q_\Omega \Lambda_\Omega Q_\Omega^\top$$
$$\hat{\Gamma}_l = Q_\Gamma \Lambda_\Gamma Q_\Gamma^\top$$

9：定义对角矩阵 Λ，并计算对角元素：

$$\Lambda^*_{ii} = n^{-1} \sum_{j=1}^{n} \left(\left(Q_{\Omega_{l-1}} \otimes Q_{\Gamma_l} \right) \nabla_{\theta_l} L_{C_j} \right)^2_i.$$

10：计算$\hat{H}_l^{-1}$

$$\hat{H}_l^{-1} = \left(Q_{\Omega_{l-1}} \otimes Q_{\Gamma_l} \right) \left(\Lambda + \lambda_l I_{d_l} \right)^{-1} \left(Q_{\Omega_{l-1}} \otimes Q_{\Gamma_l} \right)^{\mathrm{T}}$$

11: **结束循环**
12：依次将 H_l 拼接到大型对角矩阵中

$$\hat{H}_{\hat{g}}^{-1} = \begin{pmatrix} \hat{H}_1^{-1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{H}_d^{-1} \end{pmatrix}$$

其中，d 是概念预测器的卷积层编号。

13: **返回：逆海森矩阵** $\hat{H}^{-1}_{\hat{g}}$.

算法6标签预测器 f 的 EK-FAC

1: **Input:** Dataset $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^N$, original label predictor $\hat{f}$.
2：将 f 的预激活输出记为 f'，计算 $\hat{A}$:

$$A = \frac{1}{n} \cdot \sum_{i=1}^{n} \hat{g}(x_i) \cdot \hat{g}(x_i)^{\mathrm{T}}$$

3：计算 B：

$$B = \frac{1}{n} \cdot \sum_{i=1}^{n} \nabla_{f'} L_Y(\hat{f}\left(\hat{g}(x_i)\right), y_i) \cdot \nabla_{f'} L_Y(\hat{f}\left(\hat{g}(x_i)\right), y_i)^{\mathrm{T}}$$

4：对 AA 和 BB 进行特征值分解，得到Q_A、Λ_A、Q_B、Λ_B，满足

$$A = Q_A \Lambda_A Q_A^\top$$
$$B = Q_B \Lambda_B Q_B^\top$$

5：定义对角矩阵 Λ，并计算对角元素：

$$\Lambda^*_{ii} = n^{-1} \sum_{j=1}^{n} \left(\left(Q_A \otimes Q_B \right) \nabla_{\hat{f}} L_{Y_j} \right)^2_i.$$

6: 按如下方式计算$\hat{H}^{-1}$ $\hat{f}$

$$\hat{H}_{\hat{f}}^{-1} = \left(Q_A \otimes Q_B \right) \left(\Lambda + \lambda I_d \right)^{-1} \left(Q_A \otimes Q_B \right)^{\mathrm{T}}$$

7: **返回：逆海森矩阵** $\hat{H}^{-1}$ $\hat{f}$。

算法7 EK-FAC 概念-标签级CCBM

1: 输入：数据集 $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^{N}$、原始概念预测器 $\hat{f}$、标签预测器 $\hat{g}$、待移除的数据索引集 G以及阻尼参数 λ。2: 对于 $r \in G$，从 $\mathcal{D}$ 中移除第 r 个数据 (x_r, y_r, c_r)，并将新数据集定义为 $\mathcal{S}$。3: 使用算法5中的 EK-FAC 方法加速 $\hat{g}$ 的 iHVP 问题，并获得逆海森矩阵 $\hat{H}_{\hat{g}}^{-1}$ 4: 更新概念预测器 $\tilde{g}$:

$$\tilde{g} = \hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{(w,r)\in S_e} \Big(\nabla_{\hat{g}} L_{C_r} \left(\hat{g}^r(x_w), {c_w^r}' \right) - \nabla_{\hat{g}} L_{C_r} \left(\hat{g}^r(x_w), c_w^r \right) \Big).$$

5: **在算法中使用 EK-FAC 方法 6 加速** iHVP 问题，以计算 $\hat{f}$ **并得到** $\hat{H}_{\hat{f}}^{-1}$

6: 更新标签预测器 f:

$$\tilde{f} = \hat{f} + H_{\hat{f}}^{-1} \cdot \nabla_f \sum_{i=1}^{n} L_Y \left(\hat{f}\left(\hat{g}(x_i)\right), y_i \right) - H_{\hat{f}}^{-1} \cdot \nabla_f \sum_{l=1}^{n} \left(L_Y \left(\hat{f}\left(\tilde{g}(x_l)\right), y_l \right) \right).$$

7: **返回：** f、$\tilde{g}$。

算法8 EK-FAC 概念级 CCBM

1: **输入：** 数据集 $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^{n}$，原始概念预测器 f、标签预测器 $\hat{g}$ 、待移除概念索引集 M以及阻尼参数 λ。2: 对于 $r \in M$，对所有数据 $z \in \mathcal{D}$设置 $p_r = 0$ 。3: **使用算法 5 中的 EK-FAC 方法加速 $\hat{g}$ 的 iHVP 问题，并获得逆海森矩阵** $\hat{H}_{\hat{g}}^{-1}$ 4: 更新概念预测器 $\tilde{g}$:

$$\tilde{g}^* = \hat{g} - H_{\hat{g}}^{-1} \cdot \nabla_{\hat{g}} \sum_{j\notin M} \sum_{i=1}^{n} L_{C_j}(\hat{g}^j(x_i), c_i^j).$$

5: **使用算法 6 中的 EK-FAC 方法加速 $\hat{f}$ 的 iHVP 问题，并获得** $\hat{H}_{\hat{f}}^{-1}$

6: 更新标签预测器 f:

$$\tilde{f} = \hat{f} - H_{\hat{f}}^{-1} \cdot \nabla_{\hat{f}} \sum_{l=1}^{n} L_Y \left(\hat{f}\left(\tilde{g}^*(x_l)\right), y_l \right).$$

7: 将 $\tilde{g}^*$ 映射 $\tilde{g}$，方法是移除 r 在 $\tilde{g}^*$ 的最后一层中对应 $r \in M$的矩阵的第 8: **返回：** f、$\tilde{g}$。

算法 9 EK-FAC 数据级 CCBM

1: **输入：** 数据集 $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^{n}$、原始概念预测器 f和标签预测器 $\hat{g}$、一组错误数据 D_e 及其关联索引集 S_e，以及阻尼参数 λ。2: 对于 S_e中的索引 (w, r)，将第 w 个数据(x_w, y_w, c_w)的 c_w^r 更正为${c_w^r}'$ 正确标签。

3: **在算法 5 中使用 EK-FAC 方法来加速 $\hat{g}$ iHVP 问题，并求得逆海森矩阵** $\hat{H}_{\hat{g}}^{-1}$

4: 更新概念预测器 $\tilde{g}$:

$$\tilde{g} = \hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{(w,r)\in S_e} \Big(\nabla_{\hat{g}} L_{C_r} \left(\hat{g}^r(x_w), {c_w^r}' \right) - \nabla_{\hat{g}} L_{C_r} \left(\hat{g}^r(x_w), c_w^r \right) \Big).$$

5: **在算法 6 中使用 EK-FAC 方法，加速** iHVP f **问题，并得到** H^{-1} 计算f Compute A为:

$$A = H_{\hat{f}}^{-1} \cdot \sum_{i\in[n]-G} \nabla_{\hat{f}} L_Y \left(\hat{f}(\hat{g}(x_i)), y_i \right)$$

得到 f 为

$$\bar{f} = \hat{f} + A$$

6: **在算法 6 中使用 EK-FAC 方法，加速 iHVP 问题** $\bar{f}$ **并得到** $H_{\bar{f}}^{-1}$计算 B' 为

$$B' = -H_{\bar{f}}^{-1} \cdot \sum_{i\in[n]-G} \nabla_{\hat{f}} \left(L_Y(\bar{f}(\tilde{g}(x_i)), y_i) - L_Y(\bar{f}(\hat{g}(x_i)), y_i) \right)$$

更新标签预测器 f 为: $f = f + A + B'$。

7: **返回：** $\tilde{f}$、$\tilde{g}$。

10.1 影响函数的理论界

考虑数据集 $\mathcal{D}=\{(x_i,\ c_i,\ y_i\}i=1^n$，概念预测器 g 的损失函数定义为:

$$\begin{aligned}L_{\text{Total}}(\mathcal{D};g)&=\sum_{i=1}^{n}L_C(g(x_i),c_i)+\frac{\delta}{2}\cdot\|g\|^2\\&=\sum_{i=1}^{n}\sum_{j=1}^{k}L_C^j(g(x_i),c_i)+\frac{\delta}{2}\cdot\|g\|^2\\&=\sum_{i=1}^{n}\sum_{j=1}^{k}g^j(x_i)^\top\log({c_i}^j)+\frac{\delta}{2}\cdot\|g\|^2.\end{aligned}$$

从数学上看，我们有一组错误数据 D_e 及其关联索引集 $S_e\subseteq[n]\times[k]$，使得对于每个$(w,r)\in S_e$，都有$(x_w,y_w,c_w)\in D_e$，其中 c_w^r 被错误标记，而 $\tilde{c}_w^r$ 是修正后的概念标签。因此，我们的目标是估计重新训练的概念瓶颈模型。重新训练的概念预测器和标签预测器将按以下方式表示。

$$\begin{aligned}\hat{g}_e=\arg\min\Bigg[&\sum_{(i,j)\notin S_e}L_C^j\left(g(x_i),c_i\right)\\&+\sum_{(i,j)\in S_e}L_C^j\left(g(x_i),\tilde{c}_i\right)+\frac{\delta}{2}\cdot\|g\|^2\Bigg],\end{aligned}\tag{19}$$

将修正后的数据集定义为 $\mathcal{D}^*$。那么，移除错误数据 D_e 的影响后，损失函数变为

$$\begin{aligned}L^-(\mathcal{D}^*;g)=&\sum_{(i,j)\notin S_e}L_C^j\left(g(x_i),c_i\right)\\&+\sum_{(i,j)\in S_e}L_C^j\left(g(x_i),\tilde{c}_i\right)\quad+\frac{\delta}{2}\cdot\|g\|^2.\end{aligned}\tag{20}$$

假设 $\hat{}g=\arg\min L_{\text{Total}}(\mathcal{D};g)$ 是原始模型参数，而 $\hat{}g_e(\mathcal{D}^*)$ 是 $L^-(\mathcal{D}^*;g)$的极小化器，通过重新训练获得。将 $\bar{g}_e(\mathcal{D}^*)$ 记作移除错误数据 D_e 影响后得到的更新模型，该模型通过定理 11.1中的影响函数方法获得，可用于估计 $\hat{}g_e(\mathcal{D}^*)$。

为简化问题，我们着重讨论移除错误数据 D_e，而忽略将更正后的数据重新添加回去的过程。在这种情况下得到$\hat{}g_e(\mathcal{D}^*)-\bar{g}_e(\mathcal{D}^*)$ 的界后，可以用类似的方法自然地推导出重新添加更正数据时的界。为简洁起见，我们使用相同的记号。

于是，损失函数 $L^-(\mathcal{D}^*;g)$ 变为

$$\begin{aligned}L^-(\mathcal{D}^*;g)&=\sum_{(i,j)\notin S_e}L_C^j\left(g(x_i),c_i\right)+\frac{\delta}{2}\cdot\|g\|^2\\&=L_{\text{Total}}(\mathcal{D};g)-\sum_{(i,j)\in S_e}L_C^j\left(g(x_i),c_i\right)\end{aligned}\tag{21}$$

而 $\bar{g}_e(\mathcal{D}^*)$的定义变为

$$\hat{g}+H_{\hat{g}}^{-1}\cdot\sum_{(w,r)\in S_e}G_C^r(x_w,c_w;\hat{g})\tag{22}$$

其中 $H_{\hat{g}}=\nabla_{\hat{g}}^2\sum_{i,\ j}L_C^j(\hat{}g(x_i)\ \ c_i)+\delta\cdot I$ 是损失函数相对于$\hat{}g$的海森矩阵。这里 $\delta\cdot I$ 是一个较小的阻尼项，用于确保该海森矩阵的正定性。在海森矩阵中引入阻尼项，本质上等同于在初始损失函数中添加正则化项。因此，δ 也可以解释为正则化强度。

在这一部分中，我们将研究定理11.1 所给出的影响估计方法与重新训练之间的误差。我们使用参数变化作为评估指标:

$$|(\bar{g}_e-\hat{g})-(\hat{g}_e-\hat{g})|=|\bar{g}_e-\hat{g}_e|\tag{23}$$

假设 10.1. 损失函数 $L_C(x,c;g)$

$$L_C(x,c;g;j)=L_C^j(g(x),c)$$

关于 g 是凸且二次可微的，并具有正则化项 $\delta>0$。存在 $C_H\in\mathbb{R}$，使得

$$\|\nabla_g^2L_C(x,c;g_1)-\nabla_g^2L_C(x,c;g_2)\|_2\le C_H\|g_1-g_2\|_2$$

对于所有 $(x,c)\in\mathcal{D}=\{(x_i,c_i)\}_{i=1}^n$、$j\in[k]$ 以及 $g_1,g_2\in\Gamma$。

那么函数 $L'(\mathcal{D},S_e;\ g)$:

$$L'(\mathcal{D},S_e;g)=\sum_{(i,j)\in S_e}L_C^j\left(g(x_i),c_i\right)=\sum_{(i,j)\in S_e}L_C(x_i,c_i;g;j)$$

关于 g 是凸且二次可微的，并带有某种正则化。于是我们有

$$\|\nabla_g^2L'(\mathcal{D},S_e;g_1)-\nabla_g^2L'(\mathcal{D},S_e;g_2)\|_2\le|S_e|\cdot C_H\|g_1-g_2\|_2$$

对于 $g_1,g_2\in\Gamma$。

推论 10.2。

$$\|\nabla_g^2L^-(\mathcal{D}^*;g_1)-\nabla_g^2L^-(\mathcal{D}^*;g_2)\|_2\le((nk+|S_e|)\cdot C_H)\|g_1-g_2\|$$

Define $C_H^-\triangleq(nk+|S_e|)\cdot C_H$

定义 10.3。 将 $|\mathcal{D}|$ 定义为数对的数量

$$\begin{aligned}C_L'&=\|\nabla_gL'(\mathcal{D},S_e;\hat{g})\|_2,\\\sigma'_{\min}&=\text{smallest singular value of }\nabla_g^2L^-(\mathcal{D}^*;\hat{g}),\\\sigma_{\min}&=\text{smallest singular value of }\nabla_g^2L_{\text{Total}}(\mathcal{D};\hat{g}),\end{aligned}$$

基于上述推论和假设，我们得出以下定理。

定理 10.4。 我们得到实际影响与预测影响之间的误差如下:

$$\begin{aligned}&\|\hat{g}_e(\mathcal{D}^*)-\bar{g}_e(\mathcal{D}^*)\|\\\le&\frac{C_H^-{C_L'}^2}{2(\sigma'_{min}+\delta)^3}+\left|\frac{2\delta+\sigma_{min}+\sigma'_{min}}{(\delta+\sigma'_{min})\cdot(\delta+\sigma_{min})}\right|\cdot C_L'.\end{aligned}$$

证明。 我们将使用单步牛顿近似作为中间步骤。定义 $\Delta g_{Nt}(\mathcal{D}^*)$ 为

$$\Delta g_{Nt}(\mathcal{D}^*)\triangleq H_\delta^{-1}\cdot\nabla_gL'(\mathcal{D},S_e;\hat{g}),$$

其中，$H_\delta=\delta\cdot I+\nabla_g^2L^-(\mathcal{D}^*;\ \hat{}g)$ 是在$\hat{}g$ 处的正则化经验海森矩阵，但在移除错误数据的影响后重新加权。然后，$\hat{}g(\mathcal{D}^*)$ 的单步牛顿近似定义为 $g_{Nt}(\mathcal{D}^*)\triangleq\Delta g_{Nt}(\mathcal{D}^*)+\hat{}g$。

接下来，我们将把$\bar{g}_e(\mathcal{D}^*)$与 $\hat{}g_e(\mathcal{D}^*)$之间的误差分解为以下两部分:

$$\begin{aligned}\hat{g}_e(\mathcal{D}^*)-\bar{g}_e(\mathcal{D}^*)=&\underbrace{\hat{g}_e(\mathcal{D}^*)-g_{Nt}(\mathcal{D}^*)}_{\text{Err}_{\text{Nt, act}}(\mathcal{D}^*)}\\&+\underbrace{(g_{Nt}(\mathcal{D}^*)-\hat{g})-(\bar{g}_e(\mathcal{D}^*)-\hat{g})}_{\text{Err}_{\text{Nt, if}}(\mathcal{D}^*)}\end{aligned}$$

首先，在**步骤**1中，我们将推导牛顿实际误差 $\mathrm{Err}_{\mathrm{Nt,\,act}}(\mathcal{D}^*)$的界。由于 $L^-(g)$ 是参数为 $\sigma'_{\min}+\delta$ 的强凸函数，且在 $\hat{}g_e(\mathcal{D}^*)$处取得最小值，因此我们可以用在 g_{Nt}处的梯度范数来界定距离 $\|\hat{}g_e(\mathcal{D}^*)-g_{Nt}(\mathcal{D}^*)\|_2$ ：

$$\|\hat{g}_e(\mathcal{D}^*)-g_{Nt}(\mathcal{D}^*)\|_2 \leq \frac{2}{\sigma'_{\min}+\delta}\left\|\nabla_g L^-\left(g_{Nt}(\mathcal{D}^*)\right)\right\|_2 \tag{24}$$

因此，问题归结为对$\|\nabla_g L^-(g_{Nt}(\mathcal{D}^*))\|_2$进行界定。注意到 $\nabla_g L'(\hat{}g) = -\nabla_g L^-$。这是因为 $\hat{}g$ 使 L^-+L'最小化，即

$$\nabla_g L^-(\hat{g})+\nabla_g L'(\hat{g})=0.$$

回顾一下，$\Delta g_{Nt} = H_\delta^{-1}\cdot\nabla_g L'(\mathcal{D},S_e;\hat{}g) = -H_\delta^{-1}\cdot\nabla_g L^-(\mathcal{D}^*;\hat{}g)$。根据上述条件，我们可以得到 $\mathrm{Err}_{\mathrm{Nt,\,act}}(-\mathcal{D}^*)$的这一界限。

$$\begin{aligned}
&\left\|\nabla_g L^-\left(g_{Nt}(\mathcal{D}^*)\right)\right\|_2\\
=&\left\|\nabla_g L^-\left(\hat{g}+\Delta g_{Nt}(\mathcal{D}^*)\right)\right\|_2\\
=&\Big\|\nabla_g L^-\left(\hat{g}+\Delta g_{N_t}(\mathcal{D}^*)\right)-\nabla_g L^-\left(\hat{g}\right)\\
&\quad -\nabla_g^2 L^-\left(\hat{g}\right)\cdot\quad \Delta g_{N_t}(\mathcal{D}^*)\Big\|_2\\
=&\left\|\int_0^1\left(\nabla_g^2 L^-\left(\hat{g}+t\cdot\Delta g_{Nt}(\mathcal{D}^*)\right)\right.\right.\\
&\quad \left.\left.-\nabla_g^2 L^-\left(\hat{g}\right)\right)\Delta g_{Nt}(\mathcal{D}^*)\,dt\right\|_2\\
\leq&\frac{C_H^-}{2}\left\|\Delta g_{Nt}(\mathcal{D}^*)\right\|_2^2\\
=&\frac{C_H^-}{2}\left\|\left[\nabla_g^2 L^-(\hat{g})\right]^{-1}\nabla_g L^-(\hat{g})\right\|_2^2\\
\leq&\frac{C_H^-}{2(\sigma'_{\min}+\delta)^2}\left\|\nabla_g L^-(\hat{g})\right\|_2^2\\
=&\frac{C_H^-}{2(\sigma'_{\min}+\delta)^2}\left\|\nabla_g L'(\hat{g})\right\|_2^2\\
\leq&\frac{C_H^-{C'_L}^2}{2(\sigma'_{\min}+\delta)^2}.
\end{aligned} \tag{25}$$

现在我们进入**步骤** 2 ，以界定 $\mathrm{Err}_{\mathrm{Nt,\,if}}(-\mathcal{D}^*)$；我们将界定牛顿法与我们的 CCBM方法之间的参数变化差异。

$$\begin{aligned}
&\left\|\left(g_{Nt}(\mathcal{D}^*)-\hat{g}\right)-\left(\bar{g}_e(\mathcal{D}^*)-\hat{g}\right)\right\|\\
=&\Big\|\Big[\left(\delta\cdot I+\nabla_g^2 L^-\left(\hat{g}\right)\right)^{-1}+\left(\delta\cdot I+\nabla_g^2 L_{\mathrm{Total}}\left(\hat{g}\right)\right)^{-1}\Big]\\
&\cdot\nabla_g L'(\mathcal{D},S_e;\hat{g})\Big\|
\end{aligned}$$

为简化起见，以下替换中我们使用矩阵 A, B ：

$$\begin{aligned}
A&=\delta\cdot I+\nabla_g^2 L^-\left(\hat{g}\right)\\
B&=\delta\cdot I+\nabla_g^2 L_{\mathrm{Total}}\left(\hat{g}\right)
\end{aligned}$$

并且 A 和 B 是具有以下性质的正定矩阵

$$\begin{aligned}
\delta+\sigma'_{\min}&\prec A\prec\delta+\sigma'_{\max}\\
\delta+\sigma_{\min}&\prec B\prec\delta+\sigma_{\max}
\end{aligned}$$

因此，我们有

$$\begin{aligned}
&\left\|\left(g_{Nt}(\mathcal{D}^*)-\hat{g}\right)-\left(\bar{g}_e(\mathcal{D}^*)-\hat{g}\right)\right\|\\
=&\left\|\left(A^{-1}+B^{-1}\right)\cdot\nabla_g L^-(\mathcal{D}^*;\hat{g})\right\|\\
\leq&\left\|A^{-1}+B^{-1}\right\|\cdot\left\|\nabla_g L^-(\mathcal{D}^*;\hat{g})\right\|\\
\leq&\left|\frac{2\delta+\sigma_{\min}+\sigma'_{\min}}{(\delta+\sigma'_{\min})\cdot(\delta+\sigma_{\min})}\right|\cdot\left\|\nabla_g L^-(\mathcal{D}^*;\hat{g})\right\|\\
\leq&\left|\frac{2\delta+\sigma_{\min}+\sigma'_{\min}}{(\delta+\sigma'_{\min})\cdot(\delta+\sigma_{\min})}\right|\cdot C'_L
\end{aligned} \tag{26}$$

结合方程 66、67 和 68中的步骤 I 和步骤 II 的结论，我们可以得到实际影响与预测影响之间的误差如下：

$$\begin{aligned}
&\left\|\hat{g}_e(\mathcal{D}^*)-\bar{g}_e(\mathcal{D}^*)\right\|\\
\leq&\frac{C_H^-{C'_L}^2}{2(\sigma'_{\min}+\delta)^3}+\left|\frac{2\delta+\sigma_{\min}+\sigma'_{\min}}{(\delta+\sigma'_{\min})\cdot(\delta+\sigma_{\min})}\right|\cdot C'_L.
\end{aligned}$$

□

注释 10.5。 定理10.4 揭示了关于影响函数估计的一个重要发现：估计误差与正则化参数 δ ($\mathcal{O}(1/\delta)$) 成反比，表明增强正则化能够提高近似精度。

注释 10.6。 在概念瓶颈模型中，重新训练是处理移除训练数据点最准确的方法。对于概念预测器，我们推导了基于影响函数的近似的理论误差界。然而，标签预测器则有所不同。作为单层线性模型，重新训练标签预测器的计算成本较低。但其输入依赖于概念预测器，因此理论分析颇具挑战，原因如下：(1) 输入依赖：概念预测器的变化会影响标签预测器的输入，使两者的更新相互耦合。(2) 误差传播：概念预测器产生的误差会传播到标签预测器，导致复杂的相互作用。鉴于重新训练标签预测器的成本较低，直接重新训练更为实用且准确。因此，我们将理论分析重点放在概念预测器上。

11 概念标签级影响的概念-标-签影响的证明我们有一组错误数据 D_e 及其关联索引集 $S_e \subseteq [n] \times [k]$，对于每个 $(w,\ r) \in S_e$，都有 $(x_w,\ y_w,\ c_w) \in D_e$，其中 c_w^r 被错误标记，而 $\tilde{c}_w^r$ 是其修正后的概念标签。因此，我们的目标是在不重新训练的情况下近似得到新的概念瓶颈模型。

ˆ 11.0.0.1 证明概要：我们的目标是将 ˆg 和 f 编辑为 ˆˆ g_e 和 f_e。(i) 首先，我们引入新的参数 ˆ$g_{\epsilon,\ e}$，使其在带有小扰动 ϵ 的修改后的损失函数下达到最小值。(ii) 然后，我们围绕 ˆg 执行牛顿步，并获得 ˆg_e 的估计值。(iii) 接着，我们考虑在一个数据点 $(x_{i_c},\ y_{i_c},\ c_{i_c})$ 处改变概念预测器，并重新训练模型以获得新的标签预测器 f_{i_c}，ˆ 从而获得 f_{i_c} 的近似值。(iv) 接下来，我们遍历 i_c（取值为 1、2、$\cdots$、n），将所有方程相加，并围绕 f 执行 ˆˆ 牛顿步，以获得 f_e 的近似值。(v)ˆ 最后，我们将 ˆg 的估计值代入 f_e 的方程中，以获得最终近似值。

定理 11.1。由下式定义的重新训练的概念预测器 ˆg_e

$$\hat{g}_e = \arg\min \left[\sum_{(i,j)\notin S_e} L_C\left(g^j(x_i), c_i^j\right) + \sum_{(i,j)\in S_e} L_C\left(g^j(x_i), \tilde{c}_i^j\right) \right], \tag{27}$$

可近似为：

$$\hat{g}_e \approx \bar{g}_e \triangleq \hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{(w,r)\in S_e} \left(\nabla_{\hat{g}} L_C\left(\hat{g}^r(x_w), \tilde{c}_w^r\right) - \nabla_{\hat{g}} L_C\left(\hat{g}^r(x_w), c_w^r\right) \right), \tag{28}$$

其中 $H_{\hat{g}} = \nabla_{\hat{g}}^2 \sum_{i,j} L_C(\hat{}g^j(x_i)\ c_i^j)$ 是损失函数关于 ˆg 的海森矩阵。

证明。对于索引 $(w,r) \in S_e$，它表示第 w 个数据的第 r 个概念有误，我们将该概念 c_w^r 更正为 $\tilde{c}_w^r$。将 ˆg_e 改写为

$$\hat{g}_e = \arg\min \left[\sum_{i,j} L_C\left(g^j(x_i), c_i^j\right) + \sum_{(w,r)\in S_e} L_C\left(g^r(x_w), \tilde{c}_w^r\right) - \sum_{(w,r)\in S_e} L_C\left(g^r(x_w), c_w^r\right) \right]. \tag{29}$$

为近似这一影响，将新参数 ˆ$g_{\epsilon,e}$ 定义为

$$\hat{g}_{\epsilon,e} \triangleq \arg\min \left[\sum_{i,j} L_C\left(g^j(x_i), c_i^j\right) + \sum_{(w,r)\in S_e} \epsilon \cdot L_C\left(g^r(x_w), \tilde{c}_w^r\right) - \sum_{(w,r)\in S_e} \epsilon \cdot L_C\left(g^r(x_w), c_w^r\right) \right]. \tag{30}$$

然后，由于 ˆ$g_{\epsilon,\ e}$ 使方程30达到最小值，因此我们有

$$\nabla_{\hat{g}} \sum_{i,j} L_C\left(\hat{g}_{\epsilon,e}^j(x_i), c_i^j\right) + \sum_{(w,r)\in S_e} \epsilon \cdot \nabla_{\hat{g}} L_C\left(\hat{g}_{\epsilon,e}^r(x_w), \tilde{c}_w^r\right) - \sum_{(w,r)\in S_e} \epsilon \cdot \nabla_{\hat{g}} L_C\left(\hat{g}_{\epsilon,e}^r(x_w), c_w^r\right) = 0.$$

在 ˆg 处对上述方程进行泰勒展开，

$$\begin{aligned} &\nabla_{\hat{g}} \sum_{i,j} L_C\left(\hat{g}^j(x_i), c_i^j\right) + \sum_{(w,r)\in S_e} \epsilon \cdot \nabla_{\hat{g}} L_C\left(\hat{g}^r(x_w), \tilde{c}_w^r\right) \\ &- \sum_{(w,r)\in S_e} \epsilon \cdot \nabla_{\hat{g}} L_C\left(\hat{g}^r(x_w), c_w^r\right) \\ &+ \nabla_{\hat{g}}^2 \sum_{i,j} L_C\left(\hat{g}^j(x_i), c_i^j\right) \cdot (\hat{g}_{\epsilon,e} - \hat{g}) \approx 0. \end{aligned} \tag{31}$$

根据方程27，方程31的第一项等于0。因此有

$$\begin{aligned} \hat{g}_{\epsilon,e} - \hat{g} = -\sum_{(w,r)\in S_e} \epsilon \cdot H_{\hat{g}}^{-1} \cdot \left(\nabla_{\hat{g}} L_C\left(\hat{g}^r(x_w), \tilde{c}_w^r\right)\right. \\ \left. -\nabla_{\hat{g}} L_C\left(\hat{g}^r(x_w), c_w^r\right)\right), \end{aligned}$$

其中

$$H_{\hat{g}} = \nabla_{\hat{g}}^2 \sum_{i,j} L_C\left(\hat{g}^j(x_i), c_i^j\right).$$

然后，我们在 ˆg 附近进行一次牛顿步，得到

$$\hat{g}_e \approx \bar{g}_e \triangleq \hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{(w,r)\in S_e} \left(\nabla_{\hat{g}} L_C\left(\hat{g}^r(x_w), \tilde{c}_w^r\right) - \nabla_{\hat{g}} L_C\left(\hat{g}^r(x_w), c_w^r\right) \right). \tag{32}$$

□

ˆ**定理 11.2.** 由以下定义的重新训练的标签预测器 f_e

$$\hat{f}_e = \arg\min \left[\sum_{i=1}^{n} L_Y\left(f\left(\hat{g}_e\left(x_i\right)\right), y_i\right) \right]$$

可以近似为：

$$\begin{aligned} \hat{f}_e \approx \bar{f}_e = \hat{f} + H_{\hat{f}}^{-1} \cdot \nabla_f \sum_{i=1}^{n} L_{Y_i}\left(\hat{f}, \hat{g}\right) \\ - H_{\hat{f}}^{-1} \cdot \nabla_f \sum_{i=1}^{n} L_{Y_i}\left(\hat{f}, \bar{g}_e\right), \end{aligned}$$

其中 $H_{\hat{}f} = \nabla_{\hat{f}}^2 \sum_{i=1}^{n} L_{Y_i}(\hat{}f, \hat{}g)$ 是损失函数关于 ˆ f，$L_{Y_i}(\hat{}f, \hat{}g) \triangleq L_Y(\hat{}f(\hat{}g(x_i))\ y_i)$，且 $\bar{g}_e$ 在定理11.1中给出。

证明。现在我们来推导朝向 f_e 的编辑后的标签预测器 ˆ。

首先，我们只考虑在一个数据点处改变概念预测器 $(x_{i_c}, y_{i_c}, c_{i_c})$，并重新训练模型以获得 ˆ一个新的标签预测器 f_{i_c}。

$$\begin{aligned} \hat{f}_{i_c} = \arg\min \left[\sum_{i=1}^{n} L_Y\left(f\left(\hat{g}\left(x_i\right)\right), y_i\right) + L_Y\left(f\left(\hat{g}_e\left(x_{i_c}\right)\right), y_{i_c}\right) \right. \\ \left. - L_Y\left(f\left(\hat{g}\left(x_{i_c}\right)\right), y_{i_c}\right) \right]. \end{aligned}$$

我们将上述方程改写如下：

$$\begin{aligned} \hat{f}_{i_c} = \arg\min \left[\sum_{i=1}^{n} L_Y\left(f\left(\hat{g}\left(x_i\right)\right), y_i\right) + L_Y\left(f\left(\hat{g}_e\left(x_{i_c}\right)\right), y_{i_c}\right) \right. \\ \left. - L_Y\left(f\left(\hat{g}\left(x_{i_c}\right)\right), y_{i_c}\right) \right]. \end{aligned}$$

ˆ我们定义f_{ϵ,i_c}为:

$$\hat{f}_{\epsilon,i_c} = \arg\min\Bigg[\sum_{i=1}^{n} L_Y\left(f\left(\hat{g}\left(x_i\right)\right), y_i\right) + \epsilon \cdot L_Y\left(f\left(\hat{g}_e\left(x_{i_c}\right)\right), y_{i_c}\right) - \epsilon \cdot L_Y\left(f\left(\hat{g}\left(x_{i_c}\right)\right), y_{i_c}\right)\Bigg].$$

对上方方程两边关于 f 求导，可得

$$\begin{aligned}\nabla_{\hat{f}} \sum_{i=1}^{n} & L_Y\left(\hat{f}_{\epsilon,i_c}\left(\hat{g}(x_i)\right), y_i\right) \\ & + \epsilon \cdot \nabla_{\hat{f}} L_Y\left(\hat{f}_{\epsilon,i_c}\left(\hat{g}_e(x_{i_c})\right), y_{i_c}\right) \\ & - \epsilon \cdot \nabla_{\hat{f}} L_Y\left(\hat{f}_{\epsilon,i_c}\left(\hat{g}(x_{i_c})\right), y_{i_c}\right) \\ & = 0\end{aligned}$$

ˆ在 f 处对上方方程进行泰勒展开，

$$\begin{aligned}&\nabla_{\hat{f}} \sum_{i=1}^{n} L_Y\left(\hat{f}\left(\hat{g}\left(x_i\right)\right), y_i\right) + \epsilon \cdot \nabla_{\hat{f}} L_Y\left(\hat{f}\left(\hat{g}_e\left(x_{i_c}\right)\right), y_{i_c}\right) \\ &- \epsilon \cdot \nabla_{\hat{f}} L_Y\left(\hat{f}\left(\hat{g}\left(x_{i_c}\right)\right), y_{i_c}\right) \\ &+ \nabla_{\hat{f}}^2 \sum_{i=1}^{n} L_Y\left(\hat{f}\left(\hat{g}\left(x_i\right)\right), y_i\right) \cdot \left(\hat{f}_{\epsilon,i_c} - \hat{f}\right) = 0\end{aligned}$$

于是可得

$$\hat{f}_{\epsilon,i_c} - \hat{f} \approx -\epsilon \cdot H_{\hat{f}}^{-1} \cdot \nabla_f \left(L_Y\left(\hat{f}\left(\hat{g}_e\left(x_{i_c}\right)\right), y_{i_c}\right)\right.$$

其中 $H^{-1}{}_{\hat{f}} = \nabla_{\hat{f}}^2 \sum_{i=1}^{n} L_Y\left(\hat{f}\left(\hat{g}\left(x_{i_c}\right)\right), y_{i_c}\right)\Big), -L_Y\left(\hat{f}\left(\hat{g}\left(x_{i_c}\right)\right), y_{i_c}\right)$ ˆ$f(\hat{g}(x_i))$ $y_i\Big)$。

令 i_c 依次取 1、2、$\cdots$、n，并将所有方程相加，可得:

$$\begin{aligned}\hat{f}_{\epsilon,e} - \hat{f} \approx -\epsilon \cdot H_{\hat{f}}^{-1} \cdot \sum_{i=1}^{n} \nabla_f & \left(L_Y\left(\hat{f}\left(\hat{g}_e\left(x_i\right)\right), y_i\right)\right. \\ & \left. -L_Y\left(\hat{f}\left(\hat{g}\left(x_i\right)\right), y_i\right)\right).\end{aligned}$$

ˆ围绕 f 执行牛顿步，我们得到

$$\begin{aligned}\hat{f}_e \approx \hat{f} - H_{\hat{f}}^{-1} \cdot \sum_{i=1}^{n} \nabla_f & \left(L_Y\left(\hat{f}\left(\hat{g}_e\left(x_i\right)\right), y_i\right)\right. \\ & \left. -L_Y\left(\hat{f}\left(\hat{g}\left(x_i\right)\right), y_i\right)\right).\end{aligned} \tag{33}$$

将编辑后的32 中的 g 代入方程33，得到

$$\begin{aligned}\hat{f}_e &\approx \hat{f} - H_{\hat{f}}^{-1} \cdot \sum_{i=1}^{n} \nabla_f \Big(L_Y\left(\hat{f}\left(\bar{g}_e(x_i)\right), y_i\right) \\ &\quad - L_Y\left(\hat{f}\left(\hat{g}(x_i)\right), y_i\right)\Big) \\ &= \hat{f} - H_{\hat{f}}^{-1} \cdot \sum_{i=1}^{n} \nabla_f \left(L_{Y_i}(\hat{f}, \bar{g}_e) - L_{Y_i}(\hat{f}, \hat{g})\right) \\ &= \hat{f} + H_{\hat{f}}^{-1} \cdot \nabla_f \sum_{i=1}^{n} L_{Y_i}(\hat{f}, \hat{g}) \\ &\quad - H_{\hat{f}}^{-1} \cdot \nabla_f \sum_{i=1}^{n} L_{Y_i}(\hat{f}, \bar{g}_e) \\ &\triangleq \bar{f}_e.\end{aligned}$$

□

11.1 影响函数的理论界

考虑数据集 $\mathcal{D}=\{(x_i,\ c_i,\ y_i\}_{i=1}^n$，概念预测器 g 的损失函数定义如下：

$$\begin{aligned}L_{\text{Total}}(\mathcal{D};g)&=\sum_{i=1}^{n}L_C(g(x_i),c_i)+\frac{\delta}{2}\cdot\|g\|^2\\&=\sum_{i=1}^{n}\sum_{j=1}^{k}L_C^j(g(x_i),c_i^j)+\frac{\delta}{2}\cdot\|g\|^2\\&=\sum_{i=1}^{n}\sum_{j=1}^{k}g^j(x_i)^\top\log(c_i^j)+\frac{\delta}{2}\cdot\|g\|^2.\end{aligned}$$

从数学上讲，我们有一组需要移除的错误概念，记为 p_r for $r\in M$。于是，重新训练的概念预测器变为

$$\hat{g}_{-p_M}=\arg\min_{g'}\sum_{j\notin M}\sum_{i=1}^{n}L_C^j(g'(x_i),c_i)+\frac{\delta}{2}\cdot\|g\|^2.$$

我们将其映射为 $\hat{}g^*_{-p_M}$，其中 $\hat{}g_{-p_M}$ to $\hat{}g^*_{-p_M}\triangleq \mathrm{P}(\hat{}g_{-p_M})$；它与 $\hat{}g$ 具有相同数量的参数，并且对于所有 $j\in[d_i]-M$，具有相同的预测概念 $\hat{}g^*_{-p_M}(j)$ as $\hat{}g_{-p_M}(j)$。我们通过在 $\hat{}g_{-p_M}$ 最后一层矩阵的 r-th 行中插入一个零行向量来实现这一效果，for $r\in M$。因此，我们可以看出，映射 P是一一对应的。此外，假设 $\hat{}g$ 的参数空间为 T，而 $\hat{}g^*_{-p_M}$, T_0的参数空间是 T的子集。注意到，$\hat{}g^*_{-p_M}$ 是以下目标函数的最优模型：

$$\hat{g}^*_{-p_M}=\arg\min_{g'\in T_0}\sum_{j\notin M}\sum_{i=1}^{n}L_C^j(g'(x_i),c_i)+\frac{\delta}{2}\cdot\|g\|^2.$$

移除错误概念影响后的损失函数为

$$\begin{aligned}L^-(\mathcal{D};g)&=\sum_{j\notin M}\sum_{i=1}^{n}L_C^j(g'(x_i),c_i)+\frac{\delta}{2}\cdot\|g\|^2\\&=L_{\text{Total}}(\mathcal{D};g)-\sum_{j\in M}\sum_{i=1}^{n}L_C^j\left(g(x_i),c_i\right).\end{aligned}\tag{34}$$

假设 $\hat{}g=\arg\min\ L_{\text{Total}}(\mathcal{D};g)$ 为原始模型参数。$\hat{}g_{-p_M}(\mathcal{D})$ 和 $\hat{}g^*_{-p_M}(\mathcal{D})$ 是通过在不同参数空间中重新训练得到的 $L^-(\mathcal{D};g)$ 的极小化器。$\hat{}g^*_{-p_M}(\mathcal{D})$ 与原始模型具有相同的维数。由于 $\hat{}g_{-p_M}(\mathcal{D})$ 和 $\hat{}g^*_{-p_M}(\mathcal{D})$ 在给定相同输入时会产生相同输出，为了简化证明，我们将 $\hat{}g^*_{-p_M}(\mathcal{D})$ 作为重新训练后的模型。

记 $\bar{g}_{-p_M}$ 为移除错误概念影响后通过定理12.3中的影响函数方法得到的更新模型，它是$\hat{}g^*_{-p_M}(\mathcal{D})$的估计。

$$\bar{g}_{-p_M}(\mathcal{D})\triangleq\hat{g}-H_{\hat{g}}^{-1}\cdot\sum_{j\notin M}\sum_{i=1}^{n}G_C^j(x_i,c_i;\hat{g}),$$

在本部分中，我们将研究由定理 12.3 方法给出的影响估计值与$\hat{}g^*_{-p_M}(\mathcal{D})$之间的误差。我们使用参数变化作为评估指标：

$$\left|(\bar{g}_{-p_M}-\hat{g})-(\hat{g}^*_{-p_M}-\hat{g})\right|=\left|\bar{g}_{-p_M}-\hat{g}^*_{-p_M}\right|\tag{35}$$

假设 11.3. 损失 $L_C(x,c;g;\ j)$

$$L_C(\mathcal{D};g;j)=\sum_{i=1}^{n}L_C^j(g(x_i),c_i).$$

在 g 上是凸且二次可微的，并具有正则化项 $\delta>0$。存在 $C_H\in\mathbb{R}$ 使得

$$\|\nabla_g^2L_C(\mathcal{D};g_1;j)-\nabla_g^2L_C(\mathcal{D};g_2;j)\|_2\le C_H\|g_1-g_2\|_2$$

对于所有 $j\in[k]$ 和 $g_1,\ g_2\in\Gamma$。

定义 11.4。

$$C'_L=\max_j\|\nabla_gL_C(\mathcal{D};\hat{g};j)\|_2,$$

$$\sigma'_{\min}=\text{smallest singular value of }\nabla_g^2L^-(\mathcal{D};\hat{g}),$$

$$\sigma_{\min}=\text{smallest singular value of }\nabla_g^2L_{\text{Total}}(\mathcal{D};\hat{g}),$$

$$L'(\mathcal{D},M;g)=\sum_{j\in M}L_C(\mathcal{D};g;j)\tag{36}$$

推论 11.5。

$$L^-(\mathcal{D};g)=L_{Total}(\mathcal{D};g)-L'(\mathcal{D},M;g)\tag{37}$$

$$\|\nabla_g^2L^-(\mathcal{D};\ g_1)-\nabla_g^2L^-(\mathcal{D};\ g_2)\|_2\le((k+|M|)\ \cdot C_H)\ \|g_1-g_2\|$$

Define $C_H^-\triangleq(k+|M|)\cdot C_H$

基于上述推论和假设，我们推导出以下定理。

定理 11.6。 我们得到实际影响与预测影响之间的误差如下：

$$\begin{aligned}&\left\|\hat{g}^*_{-p_M}(\mathcal{D})-\bar{g}_{-p_M}(\mathcal{D})\right\|\\\le&\frac{C_H^-C'_L|M|^2}{2(\sigma'_{min}+\delta)^3}+\left|\frac{2\delta+\sigma_{min}+\sigma'_{min}}{(\delta+\sigma'_{min})\cdot(\delta+\sigma_{min})}\right|\cdot C'_L|M|.\end{aligned}$$

证明。 我们将使用单步牛顿近似作为中间步骤。定义 $\Delta g_{Nt}(\mathcal{D})$ 为

$$\Delta g_{Nt}(\mathcal{D})\triangleq H_\delta^{-1}\cdot\nabla_gL'(\mathcal{D},M;\hat{g}),$$

其中，$H_\delta=\delta\cdot I+\nabla_g^2L^-(\mathcal{D};\hat{}g)$ 是在 $\hat{}g$ 处的正则化经验海森矩阵，但在移除错误数据的影响后重新加权。于是，$\hat{}g^*_{-p_M}(\mathcal{D})$ 的单步牛顿近似定义为 $g_{Nt}(\mathcal{D})\triangleq\Delta g_{Nt}(\mathcal{D})+\hat{}g$。

下文中，我们将把 $\bar{g}_{-p_M}(\mathcal{D})$ 与 $\hat{}g^*_{-p_M}(\mathcal{D})$ 之间的误差分解为以下两部分：

$$\begin{aligned}\hat{g}^*_{-p_M}(\mathcal{D})-\bar{g}_{-p_M}(\mathcal{D})=&\underbrace{\hat{g}^*_{-p_M}(\mathcal{D})-g_{Nt}(\mathcal{D})}_{\text{Err}_{\text{Nt, act}}(\mathcal{D})}\\&+\underbrace{(g_{Nt}(\mathcal{D})-\hat{g})-(\bar{g}_{-p_M}(\mathcal{D})-\hat{g})}_{\text{Err}_{\text{Nt, if}}(\mathcal{D})}\end{aligned}$$

首先，在**步骤** 1中，我们将推导牛顿实际误差 $\text{Err}_{\text{Nt, act}}(\mathcal{D})$的界。由于 $L^-(g)$ 的参数为 $\sigma'_{\min}+\delta$，且是强凸函数，其极小值由$\hat{}g^*_{-p_M}(\mathcal{D})$取得，因此我们可以用 g_{Nt} 处梯度的范数来界定距离 $\left\|\hat{}g^*_{-p_M}(\mathcal{D})-g_{Nt}(\mathcal{D})\right\|_2$：

$$\left\|\hat{g}^*_{-p_M}(\mathcal{D})-g_{Nt}(\mathcal{D})\right\|_2\le\frac{2}{\sigma'_{\min}+\delta}\left\|\nabla_gL^-\left(g_{Nt}(\mathcal{D})\right)\right\|_2\tag{38}$$

因此，问题归结为对$\|\nabla_g L^-(g_{Nt}(\mathcal{D}))\|_2$进行界定。注意到 $\nabla_g L'(\hat{g}) = -\nabla_g L^-$。这是因为$\hat{g}$使 $L^- + L'$最小，也就是说，

$$\nabla_g L^-(\hat{g}) + \nabla_g L'(\hat{g}) = 0.$$

回顾一下，$\Delta g_{Nt} = H_\delta^{-1} \cdot \nabla_g L'(\mathcal{D}, S_e; \hat{g}) = -H_\delta^{-1} \cdot \nabla_g L^-(\mathcal{D}; \hat{g})$。给定上述条件，我们可以得到$\mathrm{Err}_{\mathrm{Nt,\ act}}(-\mathcal{D})$的这一界。

$$\begin{aligned}
&\left\|\nabla_g L^-\left(g_{Nt}(\mathcal{D})\right)\right\|_2 \\
=&\left\|\nabla_g L^-\left(\hat{g} + \Delta g_{Nt}(\mathcal{D})\right)\right\|_2 \\
=&\left\|\nabla_g L^-\left(\hat{g} + \Delta g_{N_t}(\mathcal{D})\right) - \nabla_g L^-\left(\hat{g}\right) - \nabla_g^2 L^-\left(\hat{g}\right) \cdot \Delta g_{N_t}(\mathcal{D})\right\|_2 \\
=&\left\|\int_0^1 \left(\nabla_g^2 L^-\left(\hat{g} + t \cdot \Delta g_{Nt}(\mathcal{D})\right) - \nabla_g^2 L^-\left(\hat{g}\right)\right) \Delta g_{Nt}(\mathcal{D})\, dt\right\|_2 \\
\leq&\frac{C_H^-}{2}\left\|\Delta g_{Nt}(\mathcal{D}^*)\right\|_2^2 = \frac{C_H^-}{2}\left\|\left[\nabla_g^2 L^-(\hat{g})\right]^{-1} \nabla_g L^-(\hat{g})\right\|_2^2 \\
\leq&\frac{C_H^-}{2(\sigma'_{\min} + \delta)^2}\left\|\nabla_g L^-(\hat{g})\right\|_2^2 = \frac{C_H^-}{2(\sigma'_{\min} + \delta)^2}\left\|\nabla_g L'(\hat{g})\right\|_2^2 \\
\leq&\frac{C_H^- C'_L |M|^2}{2(\sigma'_{\min} + \delta)^2}.
\end{aligned} \tag{39}$$

现在进入第**步骤2**，以界定$\mathrm{Err}_{\mathrm{Nt,\ if}}(-\mathcal{D})$；我们将界定牛顿法与我们的 CCBM方法之间参数变化的差异。

$$\begin{aligned}
&\left\|\left(g_{Nt}(\mathcal{D}) - \hat{g}\right) - \left(\bar{g}_{-p_M}(\mathcal{D}) - \hat{g}\right)\right\| \\
=&\left\|\left[\left(\delta \cdot I + \nabla_g^2 L^-(\hat{g})\right)^{-1} + \left(\delta \cdot I + \nabla_g^2 L_{\mathrm{Total}}(\hat{g})\right)^{-1}\right] \right. \\
&\left. \cdot \nabla_g L'(\mathcal{D}, S_e; \hat{g})\right\|
\end{aligned}$$

为简化起见，我们对以下替换使用矩阵 A, B：

$$\begin{aligned}
A &= \delta \cdot I + \nabla_g^2 L^-(\hat{g}) \\
B &= \delta \cdot I + \nabla_g^2 L_{\mathrm{Total}}(\hat{g})
\end{aligned}$$

并且 A 和 B 是具有以下性质的正定矩阵

$$\begin{aligned}
\delta + \sigma'_{\min} \prec A \prec \delta + \sigma'_{\max} \\
\delta + \sigma_{\min} \prec B \prec \delta + \sigma_{\max}
\end{aligned}$$

因此，我们有

$$\begin{aligned}
&\left\|\left(g_{Nt}(\mathcal{D}) - \hat{g}\right) - \left(\bar{g}_{-p_M}(\mathcal{D}) - \hat{g}\right)\right\| \\
=&\left\|\left(A^{-1} + B^{-1}\right) \cdot \nabla_g L^-(\mathcal{D}; \hat{g})\right\| \\
\leq&\left\|A^{-1} + B^{-1}\right\| \cdot \left\|\nabla_g L^-(\mathcal{D}; \hat{g})\right\| \\
\leq&\left|\frac{2\delta + \sigma_{\min} + \sigma'_{\min}}{(\delta + \sigma'_{\min}) \cdot (\delta + \sigma_{\min})}\right| \cdot \left\|\nabla_g L^-(\mathcal{D}; \hat{g})\right\| \\
\leq&\left|\frac{2\delta + \sigma_{\min} + \sigma'_{\min}}{(\delta + \sigma'_{\min}) \cdot (\delta + \sigma_{\min})}\right| \cdot C'_L |M|
\end{aligned} \tag{40}$$

结合方程66、67 和68中步骤 I 和步骤 II 的结论，我们得到实际影响与预测影响之间的误差如下：

$$\begin{aligned}
&\left\|\hat{g}^*_{-p_M}(\mathcal{D}) - \bar{g}_{-p_M}(\mathcal{D})\right\| \\
\leq&\frac{C_H^- C'_L |M|^2}{2(\sigma'_{\min} + \delta)^3} + \left|\frac{2\delta + \sigma_{\min} + \sigma'_{\min}}{(\delta + \sigma'_{\min}) \cdot (\delta + \sigma_{\min})}\right| \cdot C'_L |M|.
\end{aligned}$$

□

备注 11.7。定理11.6 揭示了影响函数估计的一项关键发现：估计误差与正则化参数 δ ($\mathcal{O}(1/\delta)$)成反比，表明增强正则化可以提高近似精度。此外，误差界随移除的概念数量 $|M|$线性增长。这意味着，移除的错误概念越多，估计误差就越大。

co

12 概念级影响

12.1 概念级影响函数的证明

我们处理删除 p_r 以用于ˆ移除概念的数据集的情况。我们的目标是估计ˆg_{-p_M}, f_{-p_M}，即在移除概念数据集上训练的概念和标签预测器。

12.1.0.1 证明概要：主要思路如下：(i) 首先，我们定义一个新的预测器 ˆ$g^*_{p_M}$，其维度与 ˆg 相同，输出也与 ˆg_{-p_M} 相同。然后推导出 ˆ$g^*_{p_M}$ 的一个近似。(ii) 接着，我们考虑将 $p_r =$ 设为 ˆ 0，而不是移除它，从而得到 $f_{p_M=0}$；根据引理 12.1，它等价于 ˆ f_{-p_M} 。我们用这个新的预测器作为替代，并对其进行估计。(iii) 接下来，我们假设只对一个数据点 (x_{i_r}, y_{i_r}, c_{i_r}) 使用更新后的概念预测器 ˆ$g^*_{p_M}$，并得到一个新的标签预测器 f_{ir}，再得到相对于 f 的 f_{ir} 的一步 ˆ ˆ 牛顿迭代近似。(iv) 最后，我们对所有数据点重复上述过程，并结合定理 12.3 中 ˆg 的估计，从而得到 f 的影响函数的闭式解。

首先，我们介绍以下引理：

引理 12.1。 对于概念瓶颈模型，如果标签ˆ预测器使用形式为 $f \cdot c$ 、输入为 c 的线性变换，那么，对于每个 $r \in M$，我们从 c 中移除第 r 个概念，并将新的输入记为 c'。将第 r 个概念设为 0，并将新的输入记为 c^0。那么，对于任意 c，我们有 ˆ $f_{-p_M} \cdot c' =$ ˆ$f_{p_M=0} \cdot c^0$。

ˆ ˆ证明。 假设 f_{-p_M} 和 $f_{p_M=0}$ 的参数空间分别为 Γ 和 Γ_0，那么存在一个满射 P: $\Gamma \to \Gamma_0$。对于任意 $\theta \in \Gamma$, $P(\theta)$ 表示针对 $r \in M$ 移除 θ 的第 r 行的操作。那么我们有:

$$P(\theta) \cdot c' = \sum_{t \notin M} \theta[j] \cdot c'[j] = \sum_t \theta[t] \mathbb{I}\{t \notin M\} c[t] = \theta \cdot c^0.$$

因此，对于第二阶段中的每个样本，两个模型的损失函数 $L_Y(\theta, c^0) = L_Y(P(\theta)\ c')$ 都相同。此外，通过公式推导，对于 $\theta' \in \Gamma_0$，对 $P^{-1}(\theta')$中的任意θ ，我们有:

$$\frac{\partial L_Y(\theta, c^0)}{\partial \theta} = \frac{\partial L_Y(P(\theta), c')}{\partial \theta'}$$

ˆ因此，如果采用相同的初始化， $f_{-p_M} \cdot c' =$ ˆ $f_{p_M=0} \cdot c^0$ 对数据集中的任意 c 均成立。 □

定理 12.2。 对于定义如下的重新训练的概念预测器 ˆg_{-p_M} :

$$\hat{g}_{-p_M} = \arg\min_{g'} \sum_{j \notin M} \sum_{i=1}^n L_C^j(g'(x_i), c_i), \tag{41}$$

我们将其映射为 ˆ$g^*_{-p_M}$，如下所示:

$$\hat{g}^*_{-p_M} = \arg\min_{g' \in T_0} \sum_{j \notin M} \sum_{i=1}^n L_C^j(g'(x_i), c_i). \tag{42}$$

我们还可以将初始的 ˆg 编辑为 ˆ$g^*_{-p_M}$，定义如下:

$$\bar{g}^*_{-p_M} \triangleq \hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{j \notin M} \sum_{i=1}^n D_C^j(x_i, c_i; \hat{g}),$$

其中 $H_{\hat{g}} = \nabla_g \sum_{j \notin M} \sum_{i=1}^n L_C^j(\hat{}g(x_i)\ c_i)$。然后，通过移除映射阶段中插入的所有零行，我们可以自然地近似 ˆ$g_{-p_M} \approx \mathrm{P}^{-1}(\hat{}g^*_{-p_M})$。

定理 12.3。 对于定义如下的重新训练的概念预测器 ˆg_{-p_M}

$$\hat{g}_{-p_M} = \arg\min_{g'} \sum_{j \notin M} \sum_{i=1}^n L_C^j(g'(x_i), c_i),$$

我们将其映射为 ˆ$g^*_{-p_M}$，如下所示:

$$\hat{g}^*_{-p_M} = \arg\min_{g' \in T_0} \sum_{j \notin M} \sum_{i=1}^n L_C^j(g'(x_i), c_i).$$

并且我们可以将初始的 ˆg 编辑为 ˆ$g^*_{-p_M}$，定义如下:

$$\bar{g}_{-p_M} \triangleq \hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{j \notin M} \sum_{i=1}^n D_C^j(x_i, c_i; \hat{g}), \tag{43}$$

其中 $H_{\hat{g}} = \nabla_g \sum_{j \notin M} \sum_{i=1}^n D_C^j(x_i,\ c_i;\ \hat{}g)$。然后，通过移除映射阶段插入的所有零行，我们可以自然地近似 ˆ $g_{-p_M} \approx \mathrm{P}^{-1}(\hat{}g^*_{-p_M})$。

证明。 在这一层级，我们考虑移除一组错误标注的概念或引入新概念的情形。由于从所有数据中移除概念后，新的概念预测器与原预测器的维度不同，我们将 $g^j(x_i)$ 记作第 j 个具有 x_i 的概念预测器，并将 c_i^j 记作数据 z_i 中的第 j 个概念。为简便起见，我们将 g 视为由 k 个概念预测器组成的集合，并将不同的列分开作为向量 $g^j(x_i)$。实际上，神经网络将 g 作为一个整体接收。

为便于比较，我们引入一种新的记法 ˆ$g^*_{-p_M}$。具体而言，我们将神经网络最后一层中 ˆg 和 ˆ$g^*_{-p_M}$ 的权重定义如下。

$$\hat{g}_{-p_M}(x) = \underbrace{\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1d_i} \\ w_{21} & w_{22} & \cdots & w_{2d_i} \\ \vdots & \vdots & \ddots & \vdots \\ w_{(k-1)1} & w_{(k-1)2} & \cdots & w_{(k-1)d_i} \end{pmatrix}}_{(k-1)\times d_i} \cdot \underbrace{\begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^{d_i} \end{pmatrix}}_{d_i \times 1}$$

$$= \underbrace{\begin{pmatrix} c_1 \\ \vdots \\ c_{r-1} \\ c_{r+1} \\ \vdots \\ c_k \end{pmatrix}}_{(k-1)\times 1}$$

$$\hat{g}^*_{-p_M}(x) = \underbrace{\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1d_i} \\ \vdots & \vdots & \ddots & \vdots \\ w_{(r-1)1} & w_{(r-1)2} & \cdots & w_{(r-1)d_i} \\ 0 & 0 & \cdots & 0 \\ w_{(r+1)1} & w_{(r+1)2} & \cdots & w_{(r+1)d_i} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k1} & w_{k2} & \cdots & w_{kd_i} \end{pmatrix}}_{k\times d_i} \cdot \underbrace{\begin{pmatrix} x^1 \\ \vdots \\ x^{r-1} \\ x^r \\ x^{r+1} \\ \vdots \\ x^{d_i} \end{pmatrix}}_{d_i\times 1}$$

$$= \underbrace{\begin{pmatrix} c_1 \\ \vdots \\ c_{r-1} \\ 0 \\ c_{r+1} \\ \vdots \\ c_k \end{pmatrix}}_{k\times 1},$$

其中， r 是索引集 M 中的一个索引。

首先，我们希望基于$\hat{}g$编辑$\hat{}g^*_{-p_M} \in T_0 = \{w_{\text{final}} = 0\} \subseteq T$，其中 w_{final} 是神经网络最后一层的参数。我们来看看$\hat{}g^*_{-p_M}$的定义：

$$\hat{g}^*_{-p_M} = \arg\min_{g'\in T_0} \sum_{j\notin M}\sum_{i=1}^{n} L_C^j(g'(x_i), c_i).$$

接下来，我们将第 r 个与概念相关的项从其余部分中分离出来，并将$\hat{}g$改写为以下形式：

$$\hat{g} = \arg\min_{g\in T} \left[\sum_{j\notin M}\sum_{i=1}^{n} L_C^j(g(x_i), c_i) + \sum_{r\in M}\sum_{i=1}^{n} L_C^r(g(x_i), c_i) \right].$$

接下来，如果将第 r 个概念部分的权重增加一个较小的 ϵ，就会得到新的参数 $\hat{}g_{\epsilon,p_M}$，下文将其简写为 $\hat{}g_\epsilon$。

$$\hat{g}_{\epsilon,p_M} \triangleq \arg\min_{g\in\mathcal{T}} \Bigg[\sum_{j\notin M}\sum_{i=1}^{n} L_C^j\left(g(x_i), c_i\right) + \epsilon \cdot \sum_{r\in M}\sum_{i=1}^{n} L_C^r\left(g(x_i), c_i\right) \Bigg].$$

显然，当 $\epsilon \to 0$ 时，$\hat{}g_\epsilon \to \hat{}g^*_{-p_M}$。我们可以根据上述定义得到最小化条件。

$$\nabla_{\hat{g}^*_{-p_M}} \sum_{j\notin M}\sum_{i=1}^{n} L_C^j(\hat{g}^*_{-p_M}(x_i), c_i) = 0. \tag{44}$$

$$\nabla_{\hat{g}_\epsilon} \sum_{j\notin M}\sum_{i=1}^{n} L_C^j(\hat{g}_\epsilon(x_i), c_i) + \epsilon \cdot \nabla_{\hat{g}_\epsilon} \sum_{r\in M}\sum_{i=1}^{n} L_C^r(\hat{g}_\epsilon(x_i), c_i) = 0.$$

对方程进行一阶泰勒展开44关于$\hat{}g_\epsilon$，可得

$$\nabla_g \sum_{j\notin M}\sum_{i=1}^{n} L_C^j\left(\hat{g}_\epsilon(x_i), c_i\right) + \nabla_g^2 \sum_{j\notin M}\sum_{i=1}^{n} L_C^j\left(\hat{g}_\epsilon(x_i), c_i\right) \cdot \left(\hat{g}^*_{-p_M} - \hat{g}_\epsilon\right) \approx 0.$$

于是有

$$\hat{g}^*_{-p_M} - \hat{g}_\epsilon = -H^{-1}_{\hat{g}_\epsilon} \cdot \nabla_g \sum_{j\notin M}\sum_{i=1}^{n} L_C^j(\hat{g}_\epsilon(x_i), c_i).$$

Where $H_{\hat{g}_\epsilon} = \nabla_g^2 \sum_{j\notin M}\sum_{i=1}^{n} L_C^j(\hat{g}_\epsilon(x_i), c_i)$.

We can see that:

When $\epsilon = 0$,

$$\hat{g}_\epsilon = \hat{g}^*_{-p_M},$$

当 $\epsilon = 1$时，$\hat{}g_\epsilon = \hat{}g$,

$$\hat{g}^*_{-p_M} - \hat{g} \approx -H^{-1}_{\hat{g}} \cdot \nabla_g \sum_{j\notin M}\sum_{i=1}^{n} L_C^j(\hat{g}(x_i), c_i),$$

其中 $H_{\hat{g}} = \nabla_g^2 \sum_{j\notin M}\sum_{i=1}^{n} L_C^j(\hat{}g(x_i)\ c_i)$。

然后，得到$\hat{}g^*_{-p_M}$的一个近似。

$$\hat{g}^*_{-p_M} \approx \hat{g} - H^{-1}_{\hat{g}} \cdot \nabla_g \sum_{j\notin M}\sum_{i=1}^{n} L_C^j(\hat{g}(x_i), c_i). \tag{45}$$

回顾梯度的定义：

$$G_C^j(x_i, c_i; \hat{g}) = L_C^j(\hat{g}(x_i), c_i)) = \hat{g}^j(x_i)^\top \cdot \log(c_i^j).$$

然后，$\hat{}g^*_{-p_M}$的近似变为

$$\bar{g}_{-p_M} \triangleq \hat{g} - H^{-1}_{\hat{g}} \cdot \sum_{j\notin M}\sum_{i=1}^{n} G_C^j(x_i, c_i; \hat{g}),$$

□

$\hat{}$定理 12.4。对于定义为 f_{-p_M} 的重新训练的标签预测器

$$\hat{f}_{-p_M} = \arg\min_{f'} \sum_{i=1}^{n} L_Y = \arg\min_{f'} \sum_{i=1}^{n} L_Y(f'(\hat{g}_{-p_M}(x_i)), y_i),$$

$\hat{}$我们可以考虑其等价形式 $f_{p_M=0}$：

$$\hat{f}_{p_M=0} = \arg\min_{f} \sum_{i=1}^{n} L_Y\left(f\left(\hat{g}^*_{-p_M}(x_i)\right), y_i\right),$$

可以通过以下方式编辑

$$\hat{f}_{p_M=0} \approx \bar{f}_{p_M=0} \triangleq \hat{f} - H^{-1}_{\hat{f}} \cdot \sum_{l=1}^{n} G_Y(x_l; \bar{g}^*_{-p_M}, \hat{f}),$$

其中 $H_{\hat{}f} = \nabla_{\hat{}f} \sum_{i=1}^{n} G_Y(x_l; \bar{g}^*_{-p_M}, \hat{}f)$ 是海森矩阵。$\bar{}$删除 $f_{p_M=0}$中 $r \in M$ 的第 r 个维度后，我们可以将其映射到 f_{-p_M}，这就是概念层级下最终编辑后的标签$\hat{}$预测器 f_{-p_M} 的近似。

$\hat{}$证明。现在，我们来近似 f_{-p_M}。注意到 f 的输入维度减少到 $k - |M|$。我们考虑在标签预测器的训练阶段，将所有数据点的 $p_r = 0$ 设为相同的值，从而得到另一个最优模型 $\hat{}f_{p_M=0}$。根据引理 12.1，我们知道对于相同的输入 $\hat{}\ \hat{}x$，$f_{p_M=0}(x) = f_{-p_M}$。而且，$f_{p_M=0}$ 和 f_{-p_M} 中对应的$\hat{}\ \hat{}$参数值相等。

现在，让我们考虑如何编辑初始 $\hat{f}$ to $\hat{}$ $f_{p_M=0}$。首先，假设我们只使用更新后的概念预测器

$\hat{}\hat{}g^*_{-p_M}$针对一条数据 $(x_{i_r}, y_{i_r}, c_{i_r})$ 并得到以下 f_{ir}，记作

$$\hat{f}_{ir} = \arg\min_f \Bigg[\sum_{i=1}^{n} L_Y\left(f\left(\hat{g}(x_i)\right), y_i\right) + L_Y\left(f\left(\hat{g}^*_{-p_M}(x_{ir})\right), y_{ir}\right) - L_Y\left(f\left(\hat{g}(x_{ir})\right), y_{ir}\right) \Bigg].$$

然后将第 i_r个数据的权重增加一些微小的 ϵ，得到以下新参数:

$$\hat{f}_{\epsilon,ir} = \arg\min_f \Bigg[\sum_{i=1}^{n} L_Y\left(f\left(\hat{g}(x_i)\right), y_i\right) + \epsilon \cdot L_Y\left(f\left(\hat{g}^*_{-p_M}(x_{ir})\right), y_{ir}\right) - \epsilon \cdot L_Y\left(f\left(\hat{g}(x_{ir})\right), y_{ir}\right) \Bigg].$$

随后推导最小化条件,

$$\begin{aligned} &\nabla_f \sum_{i=1}^{n} L_Y\left(\hat{f}_{ir}\left(\hat{g}(x_i)\right), y_i\right) \\ &\quad + \epsilon \cdot \nabla_f L_Y\left(\hat{f}_{ir}\left(\hat{g}^*_{-p_M}(x_{ir})\right), y_{ir}\right) \\ &\quad - \epsilon \cdot \nabla_f L_Y\left(\hat{f}_{ir}\left(\hat{g}(x_{ir})\right), y_{ir}\right) \\ &\quad = 0. \end{aligned}$$

$\hat{}$ $\hat{}\hat{}$如果我们展开 f 的第一项，即 $f_{ir,\epsilon} \to f(\epsilon \to 0)$，则

$$\nabla_f \sum_{i=1}^{n} L_Y\left(\hat{f}(\hat{g}(x_i)), y_i\right) + \epsilon \cdot \nabla_f L_Y\left(\hat{f}(\hat{g}^*_{-p_M}(x_{ir})), y_{ir}\right) - \epsilon \cdot \nabla_f L_Y\left(\hat{f}(\hat{g}(x_{ir})), y_{ir}\right) + \left(\nabla_f^2 \sum_{i=1}^{n} L_Y\left(\hat{f}(\hat{g}(x_i)), y_i\right)\right) \cdot (\hat{f}_{ir,\epsilon} - \hat{f}) = 0.$$

注意，$\nabla_f \sum_{i=1}^{n} L_Y(\hat{}f(\hat{}g(x_i))\ y_i) = 0$。因此我们有

$$\begin{aligned} \hat{f}_{ir,\epsilon} - \hat{f} = H_{\hat{f}}^{-1} \cdot \epsilon \\ \cdot \left(\nabla_f L_Y\left(\hat{f}\left(\hat{g}^*_{-p_M}(x_{ir})\right), y_{ir}\right) \right. \\ \left. -\nabla_f L_Y\left(\hat{f}\left(\hat{g}(x_{ir})\right), y_{ir}\right)\right). \end{aligned}$$

我们得出

$$\left.\frac{\mathrm{d}\hat{f}_{\epsilon,ir}}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = H_{\hat{f}}^{-1} \cdot \left(\nabla_{\hat{f}} L_Y(\hat{f}(\hat{g}^*_{-p_M}(x_{ir})), y_{ir}) - \nabla_{\hat{f}} L_Y(\hat{f}(\hat{g}(x_{ir})), y_{ir})\right).$$

$\hat{}$在 f 处执行一步牛顿迭代，我们得到 f_{i_r}的近似。

$$\hat{f}_{ir} \approx \hat{f} + H_{\hat{f}}^{-1} \cdot \left(\nabla_{\hat{f}} L_Y(\hat{f}(\hat{g}(x_{ir})), y_{ir}) - \nabla_{\hat{f}} L_Y(\hat{f}(\hat{g}^*_{-p_M}(x_{ir})), y_{ir})\right).$$

$\hat{}$重新考虑 f_{i_r} 的定义，对于一条数据 $(x_{i_r}, y_{i_r}, c_{i_r})$，我们使用更新后的概念预测器 $\hat{}g^*_{-p_M}$。现在，我们对所有其他数据执行这一操作并进行估计

$\hat{}f_{p_M=0}$。结合 $\hat{}$ f 定义中的最小化条件，我们得到

$$\begin{aligned} \hat{f}_{p_M=0} &\approx \hat{f} + H_{\hat{f}}^{-1} \cdot \left(\nabla_{\hat{f}} \sum_{i=1}^{n} L_Y(\hat{f}(\hat{g}(x_i)), y_i) - \nabla_{\hat{f}} \sum_{i=1}^{n} L_Y(\hat{f}(\hat{g}^*_{-p_M}(x_i)), y_i)\right) \quad (46) \\ &= \hat{f} + H_{\hat{f}}^{-1} \cdot \left(-\nabla_{\hat{f}} \sum_{i=1}^{n} L_Y(\hat{f}(\hat{g}^*_{-p_M}(x_i)), y_i)\right) \\ &= \hat{f} - H_{\hat{f}}^{-1} \sum_{l=1}^{n} \nabla_{\hat{f}} L_Y(\hat{f}(\hat{g}^*_{-p_M}(x_l)), y_l). \quad (47) \end{aligned}$$

定理12.3 给出了$\hat{}g^*_{-p_M}$的编辑版本。将其代入方程47，即可得到概念层级下最终的闭式编辑标签预测器:

$$\hat{f}_{p_M=0} \approx \bar{f}_{p_M=0} \triangleq \hat{f} - H_{\hat{f}}^{-1} \cdot \nabla_{\hat{f}} \sum_{l=1}^{n} L_{Y_l}\left(\hat{f}, \bar{g}^*_{-p_M}\right),$$

其中 $H_{\hat{}f} = \nabla^2_{\hat{f}} \sum_{i=1}^{n} L_{Y_i}(\hat{}f, \hat{}g)$ 是损失函数相对于$\hat{}$函数相对于 f的海森矩阵。回顾梯度的定义:

$$G_Y(x_l; \bar{g}^*_{-p_M}, \hat{f}) = \nabla_{\hat{f}} L_Y\left(\hat{f}\left(\bar{g}^*_{-p_M}(x_l)\right), y_l\right),$$

那么近似就变为

$$\hat{f}_{p_M=0} \approx \bar{f}_{p_M=0} \triangleq \hat{f} - H_{\hat{f}}^{-1} \cdot \sum_{l=1}^{n} G_Y(x_l; \bar{g}^*_{-p_M}, \hat{f}).$$

□

符号更新：

1. 概念标签校正：
错误的数据概念标签索引集 S_e，相应的概念预测器 $\hat{g}^{\text{corr}S_e}$。

2. 概念删除：
错误的概念索引集 M，相应的概念预测器 $\hat{g}^{\mathcal{C}_M}$。

3. 数据移除/添加：
$\mathcal{D}_{\text{rem}}$ 和 $\mathcal{D}_{\text{add}}$，对应的概念预测器 $\hat{g}^{\setminus\mathcal{D}_{\text{rem}}}$ 和 $\hat{g}^{\cup\mathcal{D}_{\text{add}}}$。

12.2 数据添加

定理12.5. 对于数据集 $\mathcal{D}=\{(x_i,y_i,c_i)\}_{i=1}^N$，给定一组待添加的新数据 $\mathcal{D}_{\text{add}}=\{\tilde{z}_s=(\tilde{x}_s,\tilde{y}_s,\tilde{c}_s)\}_{s=1}^M$。假设更新后的概念预测器 $\hat{g}^{\cup\mathcal{D}_{add}}$ 定义为

$$\begin{aligned}\hat{g}^{\cup\mathcal{D}_{add}}=\arg\min_g &\sum_{j\in[K]}\sum_{i\in[N]}L_{C_j}(g(x_i),c_i)\\&+\sum_{j\in[K]}\sum_{s\in[M]}L_{C_j}(g(\tilde{x}_S),\tilde{c}_S).\end{aligned}\tag{48}$$

定义 $L_C(\hat{g}(x_i)\ c_i)\triangleq\sum_{j=1}^K L_{C_j}(\hat{g}(x_i)\ c_i)$。那么，我们可以得到关于$\hat{g}^{\cup\mathcal{D}_{add}}$的以下近似

$$\hat{g}^{\cup\mathcal{D}_{add}}\approx\bar{g}^{\cup\mathcal{D}_{add}}\triangleq\hat{g}-H_{\hat{g}}^{-1}\cdot\sum_{s\in[M]}\nabla_g L_C(\hat{g}(\tilde{x}_s),\tilde{c}_s),\tag{49}$$

其中，$H_{\hat{g}}=\nabla_{\hat{g}}^2\sum_{i=1}^N L_C(\hat{g}(x_i)\ c_i)$ 是损失函数关于$\hat{g}$的海森矩阵。

证明. 首先，回顾$\hat{g}^{\cup\mathcal{D}_{\text{add}}}$ 的定义：

$$\hat{g}^{\cup\mathcal{D}_{\text{add}}}=\arg\min_g\left[\sum_{i=1}^N L_C(g(x_i),c_i)+\sum_{s=1}^M L_C(g(\tilde{x}_s),\tilde{c}_s)\right],$$

然后，我们将第 s 个数据的权重增加了 ϵ，并得到一个新的预测器 $\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}}$:

$$\arg\min_g\left[\sum_{i=1}^N L_C(g(x_i),c_i)+\epsilon\cdot\sum_{s=1}^M L_C(g(\tilde{x}_s),\tilde{c}_s)\right].\tag{50}$$

由于 $\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}}$使方程（85）右侧取得最小值，因此我们有

$$\nabla_g\sum_{i=1}^N L_Y(\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}}(x_i),c_i)+\epsilon\cdot\nabla_g\sum_{s=1}^M L_Y(\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}}(\tilde{x}_s),\tilde{c}_s)=0.$$

当 $\epsilon\to 0$ 时，$\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}}\to\hat{g}$。因此，我们可以针对 g 进行一阶泰勒展开，得到

$$\begin{aligned}0\approx&\nabla_g\sum_{i=1}^N L_C(\hat{g}(x_i),c_i)+\epsilon\cdot\nabla_g\sum_{s=1}^M L_C(\hat{g}(\tilde{x}_s),\tilde{c}_s)\\&+\nabla_g^2\sum_{i=1}^N L_C(\hat{g}(x_i),c_i)\cdot(\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}}-\hat{g}).\end{aligned}\tag{51}$$

回顾 $\hat{g}$ 的定义：

$$\hat{g}=\arg\min_g\sum_{i=1}^n L_Y(g(x_i),c_i),$$

于是，(86) 右侧的第一项等于 0。令$\epsilon\to 0$，则有

$$\left.\frac{\mathrm{d}\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}}}{\mathrm{d}\epsilon}\right|_{\epsilon=0}=-H_{\hat{g}}^{-1}\cdot\sum_{s=1}^M\nabla_g L_C(\hat{g}(\tilde{x}_s),\tilde{c}_s),$$

where $H_{\hat{g}}=\nabla_{\hat{g}}^2\sum_{i=1}^N L_C(\hat{g}(x_i),c_i)$.

Remember when $\epsilon\to$，$\hat{g}_\epsilon^{\cup\mathcal{D}}\to\hat{g}^{\cup\mathcal{D}_{\text{add}}}$。在 $\hat{g}$ 处执行牛顿步，便得到在概念层级上编辑原始概念预测器的方法：

$$\hat{g}^{\cup\mathcal{D}_{\text{add}}}\approx\bar{g}^{\cup\mathcal{D}_{\text{add}}}\triangleq\hat{g}-H_{\hat{g}}^{-1}\cdot\sum_{s=1}^M\nabla_g L_C(\hat{g}(\tilde{x}_s),\tilde{c}_s).$$

□

定理 12.6。 对于数据集 $\mathcal{D}=(x_i,y_i,c_i)i=1^N$，给定一组待添加的新数据 $\mathcal{D}$添加 $=\tilde{z}s=(\tilde{x}s,\tilde{y}s,\tilde{c}s)s=1^M$。假设更新后的标签预测器 $\hat{f}^{\cup\mathcal{D}_{add}}$ 定义为

$$\begin{aligned}\hat{f}^{\cup\mathcal{D}_{add}}=\arg\min_f&\sum_{i=1}^N L_Y(f(\hat{g}^{\cup\mathcal{D}_{add}}(x_i)),y_i)\\&+\sum_{s=1}^M L_Y(f(\hat{g}^{\cup\mathcal{D}_{add}}(\tilde{x}_s)),\tilde{y}_s).\end{aligned}\tag{52}$$

那么，我们可以对 $\hat{f}^{\cup\mathcal{D}_{add}}$ 作如下近似：

$$\hat{f}^{\cup\mathcal{D}_{add}}\approx\bar{f}^{\cup\mathcal{D}_{add}}\triangleq\hat{f}-H^{-1}\hat{f}\cdot\sum_{s\in[M]}\nabla_f L_Y(\hat{f}(\tilde{x}s),\tilde{y}s),\tag{53}$$

其中 $H\hat{f}=\nabla^2\hat{f}\sum_{i=1}^N L_Y(\hat{f}(x_i)\ y_i)$ 是$\hat{}$损失函数关于 f 的海森矩阵。

证明。我们可以看到，$\hat{f}^{\cup\mathcal{D}_{\text{add}}}$与$\hat{f}$之间存在巨大差距。因此，首先，我们定义 $\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$ 为

$$\arg\min_f\sum_{i=1}^N L_Y\left(f(\hat{g}(x_i)),y_i\right)+\sum_{s=1}^M L_Y\left(f(\hat{g}(\tilde{x}_s)),\tilde{y}_s\right).$$

接下来，我们定义 $\tilde{f}_\epsilon^{\cup\mathcal{D}_{\text{add}}}$ 如下，以估计 $\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$。

$$\tilde{f}_\epsilon^{\cup\mathcal{D}_{\text{add}}}=\arg\min_f\sum_{i=1}^N L_Y\left(f(\hat{g}(x_i)),y_i\right)+\epsilon\cdot\sum_{s=1}^M L_Y\left(f(\hat{g}(\tilde{x}_s)),\tilde{y}_s\right).$$

根据最小化条件，我们有

$$\begin{aligned}0=&\nabla_f\sum_{i=1}^N L_Y\left(\tilde{f}_\epsilon^{\cup\mathcal{D}_{\text{add}}}(\hat{g}(x_i)),y_i\right)\\&+\epsilon\cdot\sum_{s=1}^M\nabla_f L_Y\left(\tilde{f}_\epsilon^{\cup\mathcal{D}_{\text{add}}}(\hat{g}(\tilde{x}_s)),\tilde{y}_s\right).\end{aligned}$$

$\hat{}$在 f 处进行一阶泰勒展开，

$$\begin{aligned}&\nabla_{\hat{f}}\sum_{i=1}^N L_Y\left(\hat{f}(\hat{g}(x_i)),y_i\right)+\epsilon\cdot\nabla_{\hat{f}}\sum_{s=1}^M L_Y\left(\hat{f}(\hat{g}(\tilde{x}_s)),\tilde{y}_s\right)\\&+\nabla_{\hat{f}}^2\sum_{i=1}^N L_Y\left(\hat{f}(\hat{g}(x_i)),y_i\right)\left(\tilde{f}_\epsilon^{\cup\mathcal{D}_{\text{add}}}-\hat{f}\right)=0.\end{aligned}$$

于是，$\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$ 可近似为

$$\tilde{f}^{\cup\mathcal{D}_{\text{add}}}\approx\hat{f}-H_{\hat{f}}^{-1}\cdot\sum_{s=1}^M\nabla_{\hat{f}}L_Y\left(\hat{f}(\hat{g}(\tilde{x}_s)),\tilde{y}_s\right).\tag{54}$$

定义 $A\triangleq H^{-1}\hat{}_f\cdot\sum_{s=1}^M\nabla\hat{}_f L_Y\left(\hat{}f(\hat{}g(\tilde{x}_s))\ \tilde{y}_s\right)$则 $\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$的近似定义为

$$\bar{f}^{\cup\mathcal{D}_{\text{add}}}=\hat{f}-A\tag{55}$$

然后我们估计$\hat{}f^{\cup\mathcal{D}_{\text{add}}}$与$\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$之间的差异。回顾$\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$的定义：

$$\arg\min_f \sum_{i=1}^{N} L_Y\left(f(\hat{g}(x_i)), y_i\right) + \sum_{s=1}^{M} L_Y\left(f(\hat{g}(\tilde{x}_s)), \tilde{y}_s\right). \quad (56)$$

比较方程（87）与（91），我们仍需定义一个中间预测器$\hat{}f_{s_t}^{\cup\mathcal{D}_{\text{add}}}$为

$$\begin{aligned}\hat{f}_{s_t}^{\cup\mathcal{D}_{\text{add}}} = \arg\min_f \Bigg[& \sum_{i=1}^{N} L_Y\left(f(\hat{g}(x_i)), y_i\right) \\ & + \sum_{s=1}^{M} L_Y\left(f(\hat{g}(\tilde{x}_s)), \tilde{y}_s\right) + L_Y\left(f(\hat{g}_{\epsilon}^{\cup\mathcal{D}_{\text{add}}}(\tilde{x}_{s_t})), \tilde{y}_{s_t}\right) \\ & - L_Y\left(f(\hat{g}(\tilde{x}_{s_t})), \tilde{y}_{s_t}\right)\Bigg].\end{aligned}$$

将s_t数据的权重提高某个ϵ，我们定义$\hat{}f_{s_t\,\epsilon}^{\cup\mathcal{D}_{\text{add}}}$为

$$\begin{aligned}\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}} = \arg\min_f \Bigg[& \sum_{i=1}^{N} L_Y\left(f(\hat{g}(x_i)), y_i\right) \\ & + \sum_{s=1}^{M} L_Y\left(f(\hat{g}(\tilde{x}_s)), \tilde{y}_s\right) + \epsilon \cdot L_Y\left(f(\hat{g}^{\cup\mathcal{D}_{\text{add}}}(\tilde{x}_{s_t})), \tilde{y}_{s_t}\right) \\ & - \epsilon \cdot L_Y\left(f(\hat{g}(\tilde{x}_{s_t})), \tilde{y}_{s_t}\right)\Bigg].\end{aligned}$$

于是，根据最小化条件，我们有

$$\begin{aligned}0 = & \nabla_f \sum_{i=1}^{N} L_{Y_i}\left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right) + \sum_{s=1}^{M} L_{Y_i}\left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right) \\ & + \epsilon \cdot \nabla_f L_{Y_{st}}\left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}^{\cup\mathcal{D}_{\text{add}}}\right) - \epsilon \cdot \nabla_{\hat{f}} L_{Y_{ir}}\left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right).\end{aligned} \quad (57)$$

当$\epsilon \to 0$时，$\hat{}f_{s_t\,\epsilon}^{\cup\mathcal{D}_{\text{add}}} \to \tilde{f}^{\cup\mathcal{D}_{\text{add}}}$。然后，我们在$\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$处对方程92进行泰勒展开，得到

$$\begin{aligned}0 \approx & \nabla_f \sum_{i=1}^{N} L_{Y_i}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right) + \sum_{s=1}^{M} L_{Y_i}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right) \\ & + \epsilon \cdot \nabla_f L_{Y_{st}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}^{\cup\mathcal{D}_{\text{add}}}\right) \cdot \left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}}\right) \\ & - \epsilon \cdot \nabla_{\hat{f}} L_{Y_{ir}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right) \cdot \left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}}\right).\end{aligned}$$

整理上述方程可得

$$\begin{aligned}& \hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}} \\ & \approx -\epsilon \cdot H_{\tilde{f}}^{-1} \nabla_f \Bigg[L_{Y_{s_t}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}^{\cup\mathcal{D}_{\text{add}}}\right) \\ & \quad - L_{Y_{s_t}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\Bigg],\end{aligned}$$

其中

$$H_{\tilde{f}} = \nabla_{\hat{f}}^2 \sum_{i=1}^{N} L_{Y_i}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right) + \sum_{s=1}^{M} L_{Y_i}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right).$$

当$\epsilon = 1$时，$\hat{}f_{s_t}^{\cup\mathcal{D}_{\text{add}},-1} = \hat{}f_{s_t}^{\cup\mathcal{D}_{\text{add}}}$。然后我们在$\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$处以步长1进行一次牛顿迭代，

$$\begin{aligned}\hat{f}_{s_t}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}} \approx & -H_{\tilde{f}}^{-1} \nabla_f \left(L_{Y_{st}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}^{\cup\mathcal{D}_{\text{add}}}\right)\right. \\ & \left. - L_{Y_{st}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\right),\end{aligned}$$

遍历集合$[M]$中的s_t，我们得到

$$\begin{aligned}\hat{f}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}} \approx & -H_{\tilde{f}}^{-1} \cdot \sum_{s=1}^{M} \nabla_f \left(L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}^{\cup\mathcal{D}_{\text{add}}}\right)\right. \\ & \left. - L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\right)\end{aligned} \quad (58)$$

编辑后的版本$\hat{}g^{\cup\mathcal{D}_{\text{add}}}$已根据$\bar{g}^{\cup\mathcal{D}_{\text{add}}}$定理13.3推导得到，将这一近似代入（93），可得

$$\begin{aligned}\hat{f}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}} \approx & -H_{\tilde{f}}^{-1} \cdot \nabla_{\hat{f}} \sum_{s=1}^{M} \left[L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \bar{g}^{\cup\mathcal{D}_{\text{add}}}\right)\right. \\ & \left. - L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\right].\end{aligned} \quad (59)$$

注意，由于我们没有重新训练标签预测器，而是将其编辑为$\bar{f}^{\cup\mathcal{D}_{\text{add}}}$作为替代，因此无法直接获得$\hat{}f^{\cup\mathcal{D}_{\text{add}}}$和$H_{\tilde{f}}$。因此，我们用$\bar{f}^{\cup\mathcal{D}_{\text{add}}}$近似$\hat{}f^{\cup\mathcal{D}_{\text{add}}}$，并用$H_{\bar{f}}$近似$H_{\tilde{f}}$，其定义如下：

$$H_{\bar{f}} = \nabla_{\bar{f}}^2 \sum_{i=1}^{N} L_{Y_i}\left(\bar{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right) + \nabla_{\bar{f}}^2 \sum_{s=1}^{M} L_{Y_s}\left(\bar{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)$$

然后我们将B定义为

$$B \triangleq H_{\bar{f}}^{-1} \cdot \nabla_{\hat{f}} \left(\sum_{s=1}^{M} L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \bar{g}^{\cup\mathcal{D}_{\text{add}}}\right) - L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\right) \quad (60)$$

将式（90）与（82）结合起来，我们便可推导出最终的闭式编辑后的标签预测器：

$$\bar{f}_*^{\cup\mathcal{D}_{\text{add}}} = \bar{f}^{\cup\mathcal{D}_{\text{add}}} - B = \hat{f} - A - B.$$

□

12.3 影响函数的理论界

考虑数据集 $\mathcal{D}=\{(x_i,\ c_i,\ y_i\}_{i=1}^n$，概念预测器 g 的损失函数定义如下：

$$L_{\text{Total}}(\mathcal{D};g)=\sum_{i=1}^{n}L_C(g(x_i),c_i)+\frac{\delta}{2}\cdot\|g\|^2$$
$$=\sum_{i=1}^{n}\sum_{j=1}^{k}L_C^j(g(x_i),c_i)+\frac{\delta}{2}\cdot\|g\|^2=\sum_{i=1}^{n}\sum_{j=1}^{k}g^j(x_i)^\top\log(c_i{}^j)$$
$$+\frac{\delta}{2}\cdot\|g\|^2.$$

从数学上说，我们有一组错误数据 $z_r=(x_r,y_r,c_r)$，$r\in G$ 需要将其移除。随后，重新训练的概念预测器变为

$$\hat{g}_{-z_G}=\arg\min_g\sum_{j=1}^{k}\sum_{i\in[n]-G}L_C^j(g(x_i),c_i)+\frac{\delta}{2}\cdot\|g\|^2.$$

将新数据集定义为 $\mathcal{D}^*=\{(x_i,c_i,y_i)\}_{i\in[n]-G}$，那么移除错误数据影响后的损失函数变为

$$\begin{aligned}L^-(\mathcal{D}^*;g)&=\sum_{j=1}^{k}\sum_{i\in[n]-G}L_C^j(g(x_i),c_i)+\frac{\delta}{2}\cdot\|g\|^2\\&=L_{\text{Total}}(\mathcal{D};g)-\sum_{j=1}^{k}\sum_{i\in G}L_C^j\left(g(x_i),c_i\right).\end{aligned}\tag{61}$$

假设 $\hat{}g=\arg\min L_{\text{Total}}(\mathcal{D};g)$ 是原始模型参数。$\hat{}g_{-z_G}$ 是 $L^-(\mathcal{D}^*;g)$的极小化器。记 $\bar{g}_{-z_G}$为移除错误数据影响后得到的更新模型，该模型通过定理13.1中的影响函数方法获得，是对 $\hat{}g_{-z_G}$ 的估计。

$$\hat{g}_{-z_G}\approx\bar{g}_{-z_G}\triangleq\hat{g}+H_{\hat{g}}^{-1}\cdot\sum_{r\in G}\sum_{j=1}^{M}G_C^j(x_r,c_r;\hat{g}),\tag{62}$$

在本部分中，我们将研究定理 13.1 方法给出的影响估计与 $\hat{}g_{-z_G}$ 之间的误差。我们使用参数变化作为评估指标：

$$|(\bar{g}_{-z_G}-\hat{g})-(\hat{g}_{-z_G}-\hat{g})|=|\bar{g}_{-z_G}-\hat{g}_{-z_G}|\tag{63}$$

假设12.7。损失函数 $L_C(x,\ c;\ g;\ j)$

$$L_C(x,c;g)=\sum_{j=1}^{k}L_C^j(g(x),c).$$

关于 g 是凸且二次可微的，并具有正则化项 $\delta>0$。存在 $C_H\in\mathbb{R}$，使得

$$\|\nabla_g^2L_C(x,c;g_1)-\nabla_g^2L_C(x,c;g_2)\|_2\le C_H\|g_1-g_2\|_2$$

对于所有$(x,\ c)\in\mathcal{D}$以及 $g_1,\ g_2\in\Gamma$。

定义12.8。

$$\begin{aligned}C_L'&=\|\nabla_gL_C(\mathcal{D};\hat{g})\|_2,\\\sigma_{\min}'&=\nabla_g^2L^-(\mathcal{D};\hat{}g)\text{ 的最小奇异值},\\\sigma_{\min}&=\nabla_g^2L_{\text{Total}}(\mathcal{D};\hat{}g)\text{的最小奇异值},\\L'(\mathcal{D},G;g)&=\sum_{i\in G}L_C(x_i,c_i;g)\end{aligned}\tag{64}$$

推论 12.9。

$$L^-(\mathcal{D};g)=L_{Total}(\mathcal{D};g)-L'(\mathcal{D},G;g)\tag{65}$$
$$\|\nabla_g^2L^-(\mathcal{D};\ g_1)-\nabla_g^2L^-(\mathcal{D};\ g_2)\|_2\le((n+|G|)\ \cdot C_H)$$
$\|g_1-g_2\|$定义 $C_H^-\triangleq(n+|G|)\ \cdot C_H$

基于上述推论和假设，我们得出以下定理。

定理 12.10. 我们得到实际影响与预测影响之间的误差如下：

$$\begin{aligned}&\|\hat{g}_{-z_G}(\mathcal{D})-\bar{g}_{-z_G}(\mathcal{D})\|\\\le&\frac{C_H^-C_L'|G|^2}{2(\sigma_{min}'+\delta)^3}+\left|\frac{2\delta+\sigma_{min}+\sigma_{min}'}{(\delta+\sigma_{min}')\cdot(\delta+\sigma_{min})}\right|\cdot C_L'|G|.\end{aligned}$$

证明。 我们将使用单步牛顿近似作为中间步骤。定义 $\Delta g_{Nt}(\mathcal{D})$ 为

$$\Delta g_{Nt}(\mathcal{D})\triangleq H_\delta^{-1}\cdot\nabla_gL'(\mathcal{D},G;\hat{g}),$$

其中，$H_\delta=\delta\cdot I+\nabla_g^2L^-(\mathcal{D};\hat{}g)$ 是$\hat{}g$ 处的正则化经验海森矩阵，但在移除错误数据的影响后进行了重新加权。于是，$\hat{}g_{-z_G}(\mathcal{D})$ 的单步牛顿近似定义为 $g_{Nt}(\mathcal{D})\triangleq\Delta g_{Nt}(\mathcal{D})+\hat{}g$。

接下来，我们将$\bar{g}_{-z_G}(\mathcal{D})$ 与$\hat{}g_{-z_G}(\mathcal{D})$ 之间的误差分为以下两部分：

$$\begin{aligned}\hat{g}_{-z_G}(\mathcal{D})-\bar{g}_{-z_G}(\mathcal{D})=&\underbrace{\hat{g}_{-z_G}(\mathcal{D})-g_{Nt}(\mathcal{D})}_{\text{Err}_{\text{Nt, act}}(\mathcal{D})}\\&+\underbrace{(g_{Nt}(\mathcal{D})-\hat{g})-(\bar{g}_{-z_G}(\mathcal{D})-\hat{g})}_{\text{Err}_{\text{Nt, if}}(\mathcal{D})}\end{aligned}$$

首先，在**步骤** 1中，我们将推导牛顿实际误差 $\text{Err}_{\text{Nt, act}}(\mathcal{D})$的界。由于 $L^-(g)$ 关于参数 $\sigma_{\min}'+\delta$ 是强凸的，且在$\hat{}g_{-z_G}(\mathcal{D})$处取得最小值，因此我们可以根据 g_{Nt}处梯度的范数来界定距离 $\|\hat{}g_{-z_G}(\mathcal{D})-g_{Nt}(\mathcal{D})\|_2$：

$$\|\hat{g}_{-z_G}(\mathcal{D})-g_{Nt}(\mathcal{D})\|_2\le\frac{2}{\sigma_{\min}'+\delta}\left\|\nabla_gL^-\left(g_{Nt}(\mathcal{D})\right)\right\|_2\tag{66}$$

因此，问题归结为对$\|\nabla_gL^-\left(g_{Nt}(\mathcal{D})\right)\|_2$进行界定。注意到 $\nabla_gL'(\mathcal{D},G;\ \hat{}g)=-\nabla_gL^-$。这是因为$\hat{}g$ 使 L^-+L'最小，也就是说，

$$\nabla_gL^-(\hat{g})+\nabla_gL'(\mathcal{D},G;\hat{g})=0.$$

回顾一下 $\Delta g_{Nt}=H_\delta^{-1}\cdot\nabla_gL'(\mathcal{D},\ G;\ \hat{}g)=-H_\delta^{-1}\cdot\nabla_gL^-(\mathcal{D};\ \hat{}g)$。根据上述条件，我们可以得到 $\text{Err}_{\text{Nt, act}}(-\mathcal{D})$ 的这个界。

$$\begin{aligned}&\left\|\nabla_gL^-\left(g_{Nt}(\mathcal{D})\right)\right\|_2\\=&\left\|\nabla_gL^-\left(\hat{g}+\Delta g_{Nt}(\mathcal{D})\right)\right\|_2\\=&\left\|\nabla_gL^-\left(\hat{g}+\Delta g_{N_t}(\mathcal{D})\right)-\nabla_gL^-\left(\hat{g}\right)-\nabla_g^2L^-\left(\hat{g}\right)\cdot\Delta g_{N_t}(\mathcal{D})\right\|_2\\=&\left\|\int_0^1\left(\nabla_g^2L^-\left(\hat{g}+t\cdot\Delta g_{Nt}(\mathcal{D})\right)-\nabla_g^2L^-\left(\hat{g}\right)\right)\Delta g_{Nt}(\mathcal{D})\,dt\right\|_2\\\le&\frac{C_H^-}{2}\left\|\Delta g_{Nt}(\mathcal{D}^*)\right\|_2^2=\frac{C_H^-}{2}\left\|\left[\nabla_g^2L^-(\hat{g})\right]^{-1}\nabla_gL^-(\hat{g})\right\|_2^2\\\le&\frac{C_H^-}{2(\sigma_{\min}'+\delta)^2}\left\|\nabla_gL^-(\hat{g})\right\|_2^2=\frac{C_H^-}{2(\sigma_{\min}'+\delta)^2}\left\|\nabla_gL'(\mathcal{D},G;\hat{g})\right\|_2^2\\\le&\frac{C_H^-C_L'|G|^2}{2(\sigma_{\min}'+\delta)^2}.\end{aligned}\tag{67}$$

现在我们来到**第 2 步**，为 $\mathrm{Err}_{\mathrm{Nt,\ if}}(-\mathcal{D})$ 给出界，并将比较牛顿法与我们的 CCBM 方法在参数变化上的差异。

$$\begin{aligned}
&\|(g_{Nt}(\mathcal{D})-\hat{g})-(\bar{g}_{-z_G}(\mathcal{D})-\hat{g})\| \\
&=\Big\|\Big[\left(\delta\cdot I+\nabla_g^2 L^-\left(\hat{g}\right)\right)^{-1} \\
&\quad+\left(\delta\cdot I+\nabla_g^2 L_{\text{Total}}\left(\hat{g}\right)\right)^{-1}\Big]\cdot\nabla_g L'(\mathcal{D},G;\hat{g})\Big\|
\end{aligned}$$

为简化起见，我们用矩阵 A, B 表示以下替换：

$$\begin{aligned}
A&=\delta\cdot I+\nabla_g^2 L^-\left(\hat{g}\right)\\
B&=\delta\cdot I+\nabla_g^2 L_{\text{Total}}\left(\hat{g}\right)
\end{aligned}$$

并且 A 和 B 是具有以下性质的正定矩阵

$$\begin{aligned}
\delta+\sigma'_{\min}\prec A\prec\delta+\sigma'_{\max}\\
\delta+\sigma_{\min}\prec B\prec\delta+\sigma_{\max}
\end{aligned}$$

因此，我们有

$$\begin{aligned}
&\|(g_{Nt}(\mathcal{D})-\hat{g})-(\bar{g}_{-z_G}(\mathcal{D})-\hat{g})\| \\
=&\left\|\left(A^{-1}+B^{-1}\right)\cdot\nabla_g L^-(\mathcal{D};\hat{g})\right\| \\
\le&\left\|A^{-1}+B^{-1}\right\|\cdot\left\|\nabla_g L^-(\mathcal{D};\hat{g})\right\| \\
\le&\left|\frac{2\delta+\sigma_{\min}+\sigma'_{\min}}{(\delta+\sigma'_{\min})\cdot(\delta+\sigma_{\min})}\right|\cdot\left\|\nabla_g L^-(\mathcal{D};\hat{g})\right\| \\
\le&\left|\frac{2\delta+\sigma_{\min}+\sigma'_{\min}}{(\delta+\sigma'_{\min})\cdot(\delta+\sigma_{\min})}\right|\cdot C'_L|G|
\end{aligned}\tag{68}$$

结合方程66、67 和68中步骤 I 和步骤 II 的结论，我们得到实际影响与预测影响之间的误差如下：

$$\begin{aligned}
&\|\hat{g}_{-z_G}(\mathcal{D})-\bar{g}_{-z_G}(\mathcal{D})\| \\
\le&\frac{C_H^- C'_L|G|^2}{2(\sigma'_{\min}+\delta)^3}+\left|\frac{2\delta+\sigma_{\min}+\sigma'_{\min}}{(\delta+\sigma'_{\min})\cdot(\delta+\sigma_{\min})}\right|\cdot C'_L|G|.
\end{aligned}$$

□

注释 12.11。 误差界随移除的数据数量线性增加 $|G|$。这意味着，移除的错误数据越多，估计误差就越大。

13 数据级影响的证明

13.1 数据删除

我们讨论以下情形：对于数据集 $\mathcal{D}=\{(x_i,y_i,c_i)\}_{i=1}^n$，给定一组待移除的数据 $z_r=(x_r,y_r,c_r),\ r\in G$。$\hat{}$我们的目标是估计 $\hat{}g_{-z_G}, f_{-z_G}$，它们是在移除 $r\in G$ 后的数据集上训练得到的概念和标签预测器。

13.1.0.1 证明概要。：(i) 首先，我们估计重新训练的概念预测器 $\hat{}g_{-z_G}$。(ii) 然后，我们定义一个新的 $\tilde{}$ $\tilde{}\hat{}$标签预测器 f_{-z_G}，并估计 $f_{-z_G}-f$。(iii) 接下来，为了降低计算复杂度，使用引理方法获得 $\tilde{}\hat{}\tilde{}f_{-z_G}$ 的海森矩阵的近似。(iv) 接下来，我们计算差值 $f_{-z_G}-f_{-z_G}$，其表达式为

$$-H^{-1}_{\tilde{f}_{-z_G}}\cdot\left(\nabla_{\hat{f}}L_Y\left(\tilde{f}_{-z_G}(\hat{g}_{-z_G}(x_{i_r})),y_{i_r}\right)-\nabla_{\hat{f}}L_Y\left(\tilde{f}_{-z_G}(\hat{g}(x_{i_r})),y_{i_r}\right)\right).$$

$\hat{}$ $\hat{}$(v) 最后，我们将我们实际关心的 $f_{-z_G}-f$ 分解为 $\left(\hat{}f_{-z_G}-\tilde{f}_{-z_G}\right)+\left(\tilde{f}_{-z_G}-\hat{}f\right)$。

定理 13.1。对于数据集 $\mathcal{D}=\{(x_i,\ y_i,\ c_i)\}_{i=1}^n$，给定一组待移除的数据 $z_r=(x_r,\ y_r,\ c_r),\ r\in G$。假设更新后的概念预测器 $\hat{}g_{-z_G}$ 定义为

$$\hat{g}_{-z_G}=\arg\min_g\sum_{j\in[k]}\sum_{i\in[n]-G}L_{C_j}(\hat{g}(x_i),c_i)$$

其中 $L_C(\hat{}g(x_i)\ c_i)\triangleq\sum_{j=1}^k L_{C_j}(\hat{}g(x_i)\ c_i)$。那么，我们对 $\hat{}g_{-z_G}$有如下近似

$$\hat{g}_{-z_G}\approx\bar{g}_{-z_G}\triangleq\hat{g}+H^{-1}_{\hat{g}}\cdot\sum_{r\in G}\nabla_g L_C(\hat{g}(x_r),c_r),\tag{69}$$

其中 $H_{\hat{g}}=\nabla^2_{\hat{g}}\sum_{i,\ j}L_C(\hat{}g^j(x_i)\ c_i^j)$ 是损失函数关于 $\hat{}g$的海森矩阵。

证明。首先，我们将 $\hat{}g_{-z_G}$ 重写为

$$\hat{g}_{-z_G}=\arg\min_g\left[\sum_{i=1}^n L_C(\hat{g}(x_i),c_i)-\sum_{r\in G}L_C(g(x_r),c_r)\right],$$

然后，我们将第 r 个数据的权重增加 ϵ，得到一个新的预测器 $\hat{}g_{-z_G,\ \epsilon}$，简记为 $\hat{}g_\epsilon$:

$$\hat{g}_\epsilon\triangleq\arg\min_g\left[\sum_{i=1}^n L_C(g(x_i),c_i)-\epsilon\cdot\sum_{r\in G}L_C(g(x_r),c_r)\right].\tag{70}$$

由于 $\hat{}g_\epsilon$ <style id='6}使方程右侧70达到最小值，因此有

$$\nabla_{\hat{g}_\epsilon}\sum_{i=1}^n L_Y(\hat{g}_\epsilon(x_i),c_i)-\epsilon\cdot\nabla_{\hat{g}_\epsilon}\sum_{r\in G}L_Y(\hat{g}_\epsilon(x_r),c_r)=0.$$

当 $\epsilon\to 0$ 时，$\hat{}g_\epsilon\to\hat{}g_0$。因此，我们可以关于 $\hat{}g$ 进行一阶泰勒展开，并得到

$$\nabla_g\sum_{i=1}^n L_C(\hat{g}(x_i),c_i)-\epsilon\cdot\nabla_g\sum_{r\in G}L_C(\hat{g}(x_r),c_r)+\nabla^2_g\sum_{i=1}^n L_C(\hat{g}(x_i),c_i)\cdot(\hat{g}_\epsilon-\hat{g})\approx 0.\tag{71}$$

回顾 $\hat{}g$ 的定义：

$$\hat{g}=\arg\min_g\sum_{i=1}^n L_Y(g(x_i),c_i),$$

于是，方程 71的第一项等于 0。令 $\epsilon\to 0$，则有

$$\left.\frac{\mathrm{d}\hat{g}_\epsilon}{\mathrm{d}\epsilon}\right|_{\epsilon=0}=H^{-1}_{\hat{g}}\cdot\sum_{r\in G}\nabla_g L_C(\hat{g}(x_r),c_r),$$

其中 $H^{-1}{}_{\hat{g}}=\nabla^2_g\sum_{i=1}^n\ell(\hat{}g(x_i)\ c_i)$。

记住，当 $\epsilon\to 0$ 时，$\hat{}g_\epsilon\to\hat{}g_{-z_G}$。在 $\hat{}g$ 处执行一个牛顿步，然后我们得到在概念层级下编辑原始概念预测器的方法：

$$\hat{g}_{-z_G}\approx\bar{g}_{-z_G}\triangleq\hat{g}+H^{-1}_{\hat{g}}\cdot\sum_{r\in G}\nabla_g L_C(\hat{g}(x_r),c_r).$$

□

定理 13.2。对于数据集 $\mathcal{D}=\{(x_i,y_i,c_i)\}_{i=1}^n$，给定一组要移除的数据 $z_r=(x_r,y_r,c_r),\ r\in G$。在修订后的数据集上训练的标签预测器 $\hat{}f_{-z_G}$ 变为

$$\hat{f}_{-z_G}=\arg\min_f\sum_{i\in[n]-G}L_{Y_i}(f,\hat{g}_{-z_G}).\tag{72}$$

$\tilde{}$中间标签预测器 f_{-z_G} 定义为

$$\tilde{f}_{-z_G}=\arg\min\sum_{i\in[n]-G}L_{Y_i}(f,\hat{g}),$$

$\tilde{}$ $\hat{}$于是 $f_{-z_G}-f$ 可以近似为

$$\tilde{f}_{-z_G}-\hat{f}\approx H^{-1}_{\hat{f}}\cdot\sum_{i\in[n]-G}\nabla_{\hat{f}}L_{Y_i}(\hat{f},\hat{g})\triangleq A_G.\tag{73}$$

我们将 $\tilde{f}_{-z_G}$ 的编辑版本记为 $\bar{f}^*_{-z_G}\triangleq\hat{}f+A_G$。并且，$\hat{}\tilde{}f_{-z_G}-f_{-z_G}$ 可近似为

$$\hat{f}_{-z_G}-\tilde{f}_{-z_G}\approx-H^{-1}_{\bar{f}^*_{-z_G}}\cdot\left(\nabla_{\hat{f}}\sum_{i\in[n]-G}L_{Y_i}\left(\bar{f}^*_{-z_G},\bar{g}_{-z_G}\right)-\nabla_{\hat{f}}\sum_{i\in[n]-G}L_{Y_i}\left(\bar{f}^*_{-z_G},\hat{g}\right)\right)\triangleq B_G,\tag{74}$$

其中，$H_{\bar{f}^*_{-z_G}}=\nabla_{\bar{f}}\sum_{i\in[n]-G}L_{Y_i}\left(\bar{f}^*_{-z_G},\hat{}g\right)$ 是关于 $\bar{}f^*_{-z_G}$ 在中间数据集上的损失函数的海森矩阵。然后，最终编辑后的标签预测器 f_{-z_G} 可由下式得到

$$\bar{f}_{-z_G}=\bar{f}^*_{-z_G}+B_G=\hat{f}+A_G+B_G.\tag{75}$$

$\hat{}$证明。我们可以看到，f_{-z_G} 与 $\hat{}\tilde{}f$ 之间存在巨大差距。因此，首先，我们将 f_{-z_G} 定义为

$$\tilde{f}_{-z_G}=\arg\min_f\sum_{i=1}^n L_Y\left(f(\hat{g}(x_i)),y_i\right)-\sum_{r\in G}L_Y\left(f(\hat{g}(x_r)),y_r\right).$$

$\tilde{}$然后，我们将 $f_{\epsilon,\ -z_G}$ 定义如下，以估计 f_{-z_G}。

$$\tilde{f}_{\epsilon,-z_G}=\arg\min_f\sum_{i=1}^n L_Y\left(f(\hat{g}(x_i)),y_i\right)-\epsilon\cdot\sum_{r\in G}L_Y\left(f(\hat{g}(x_r)),y_r\right).$$

根据最小化条件，我们有

$$\nabla_{\tilde{f}} \sum_{i=1}^{n} L_Y\left(\tilde{f}_{\epsilon,-z_G}(\hat{g}(x_i)), y_i\right) - \epsilon \cdot \sum_{r\in G} \nabla_{\tilde{f}} L_Y\left(\tilde{f}_{\epsilon,-z_G}(\hat{g}(x_r)), y_r\right) = 0.$$

在 f 处进行一阶泰勒展开，

$$\nabla_{\hat{f}} \sum_{i=1}^{n} L_Y\left(\hat{f}(\hat{g}(x_i)), y_i\right) - \epsilon \cdot \nabla_{\hat{f}} \sum_{r\in G} L_Y\left(\hat{f}(\hat{g}(x_r)), y_r\right) + \nabla^2_{\hat{f}} \sum_{i=1}^{n} L_Y\left(\hat{f}(\hat{g}(x_i)), y_i\right) \cdot \left(\tilde{f}_{\epsilon,-z_G} - \hat{f}\right) = 0.$$

于是，f_{-z_G} 可近似为

$$\tilde{f}_{-z_G} \approx \hat{f} + H^{-1}_{\hat{f}} \cdot \sum_{r\in G} \nabla_{\hat{f}} L_Y\left(\hat{f}(\hat{g}(x_r)), y_r\right) \triangleq A_G. \quad (76)$$

于是，f_{-z_G} 的编辑版本定义为

$$\bar{f}^*_{-z_G} = \hat{f} + A_G \quad (77)$$

于是，我们估计 f_{-z_G} 与 f_{-z_G} 之间的差异。将 f_{-z_G} 重写为

$$\tilde{f}_{-z_G} = \arg\min_f \sum_{i\in S}^{n} L_Y\left(f(\hat{g}(x_i)), y_i\right), \quad (78)$$

其中 $S \triangleq [n] - G$。

比较方程 72和 78，我们仍需定义一个中间预测器$f_{-z_G,ir}$为

$$\hat{f}_{-z_G,ir} = \arg\min_f \left[\sum_{\substack{i\in S\\ i\neq i_r}} L_{Y_i}\left(f, \hat{g}(x_i)\right) + L_{Y_{ir}}\left(f, \hat{g}_{-z_G}\right)\right] = \arg\min_f \left[\sum_{i\in S} L_{Y_i}\left(f, \hat{g}\right) + L_{Y_{ir}}\left(f, \hat{g}_{-z_G}\right) - L_{Y_{ir}}\left(f, \hat{g}\right)\right].$$

将 i_r 数据的权重提高一些 ϵ，我们定义$f_{\epsilon,-z_G,i_r}$ 为

$$\hat{f}_{\epsilon,-z_G,ir} = \arg\min_f \left[\sum_{i\in S} L_{Y_i}(f, \hat{g}) + \epsilon \cdot L_{Y_{ir}}(f, \hat{g}_{-z_G}) - \epsilon \cdot L_{Y_{ir}}(f, \hat{g})\right].$$

在下文的证明中，我们将 $f_{\epsilon,-z_G,ir}$ 记为 f^*_ϵ。那么，根据最小化条件，我们有

$$\nabla_{\hat{f}} \sum_{i\in S} L_{Y_i}\left(\hat{f}^*_\epsilon, \hat{g}\right) + \epsilon \cdot \nabla_{\hat{f}} L_{Y_{ir}}\left(\hat{f}^*_\epsilon, \hat{g}_{-z_G}\right) - \epsilon \cdot \nabla_{\hat{f}} L_{Y_{ir}}\left(\hat{f}^*_\epsilon, \hat{g}(x_{i_r})\right). \quad (79)$$

当 $\epsilon \to 0$, $f^*_\epsilon \to f_{-z_G}$。然后，我们对 在 f_{-z_G} 方程 79 进行泰勒展开，得到

$$\nabla_{\hat{f}} \sum_{i\in S} L_{Y_i}\left(\tilde{f}_{-z_G}, \hat{g}\right) + \epsilon \cdot \nabla_{\hat{f}} L_{Y_{ir}}\left(\tilde{f}_{-z_G}, \hat{g}_{-z_G}\right) - \epsilon \cdot \nabla_{\hat{f}} L_{Y_{ir}}\left(\tilde{f}_{-z_G}, \hat{g}\right) + \nabla^2_{\hat{f}} \sum_{i\in S} L_{Y_i}\left(\tilde{f}_{-z_G}, \hat{g}\right) \cdot \left(\hat{f}^*_\epsilon - \tilde{f}_{-z_G}\right) \approx 0.$$

整理上述方程可得

$$\hat{f}^*_\epsilon - \tilde{f}_{-z_G} \approx -\epsilon \cdot H^{-1}_{\tilde{f}_{-z_G}} \cdot \left(\nabla_{\hat{f}} L_{Y_{ir}}\left(\tilde{f}_{-z_G}, \hat{g}_{-z_G}\right) - \nabla_{\hat{f}} L_{Y_{ir}}\left(\tilde{f}_{-z_G}, \hat{g}\right)\right),$$

其中 $H_{\tilde{f}_{-z_G}} = \nabla^2_{\hat{f}} \sum_{i\in S} L_{Y_i}\left(\tilde{f}_{-z_G}, \hat{g}\right)$。

当 $\epsilon = 1$时，$\hat{f}^*_\epsilon = \hat{f}_{-z_G,ir}$。然后，我们在 f_{-z_G}处以步长 1 执行一次牛顿法迭代，

$$\hat{f}_{-z_G,ir} - \tilde{f}_{-z_G} \approx -H^{-1}_{\tilde{f}_{-z_G}} \cdot \left(\nabla_{\hat{f}} L_{Y_{ir}}\left(\tilde{f}_{-z_G}, \hat{g}_{-z_G}\right) - \nabla_{\hat{f}} L_{Y_{ir}}\left(\tilde{f}_{-z_G}, \hat{g}\right)\right)$$

遍历集合 S 中的 i_r，我们得到

$$\hat{f}_{-z_G} - \tilde{f}_{-z_G} \approx -H^{-1}_{\tilde{f}_{-z_G}} \cdot \left(\nabla_{\hat{f}} \sum_{i\in S} L_{Y_i}\left(\tilde{f}_{-z_G}, \hat{g}_{-z_G}\right) - \nabla_{\hat{f}} \sum_{i\in S} L_{Y_i}\left(\tilde{f}_{-z_G}, \hat{g}\right)\right) \quad (80)$$

编辑后的$\hat{g}_{-z_G}$已推导为 $\bar{g}_{-z_G}$，根据定理13.1，将此近似代入（80），得到

$$\hat{f}_{-z_G} - \tilde{f}_{-z_G} \approx -H^{-1}_{\tilde{f}_{-z_G}} \cdot \left(\nabla_{\hat{f}} \sum_{i\in S} L_{Y_i}\left(\tilde{f}_{-z_G}, \bar{g}_{-z_G}\right) - \nabla_{\hat{f}} \sum_{i\in S} L_{Y_i}\left(\tilde{f}_{-z_G}, \hat{g}\right)\right). \quad (81)$$

注意，由于我们没有重新训练标签预测器，而是对其进行编辑，将其$\hat{}$替换为 $f^*_{-z_G}$，因此无法直接获得 f_{-z_G} 和 $H_{\tilde{f}_{-z_G}}$。因此，我们用 $f^*_{-z_G}$ 近似 f_{-z_G}，并用 $H_{\bar{f}^*_{-z_G}}$ 近似 $H_{\tilde{f}_{-z_G}}$，其定义如下:

$$H_{\bar{f}^*_{-z_G}} = \nabla^2_{\hat{f}} \sum_{i\in S} L_{Y_i}\left(\bar{f}^*_{-z_G}, \hat{g}\right)$$

然后我们将 B_G 定义为

$$B_G \triangleq -H^{-1}_{\bar{f}^*_{-z_G}} \cdot \left(\nabla_{\hat{f}} \sum_{i\in S} L_{Y_i}\left(\bar{f}^*_{-z_G}, \bar{g}_{-z_G}\right) - \nabla_{\hat{f}} \sum_{i\in S} L_{Y_i}\left(\bar{f}^*_{-z_G}, \hat{g}\right)\right) \quad (82)$$

结合（90）和（82），我们推导出最终的闭式编辑后的标签预测器为

$$\bar{f}_{-z_G} = \bar{f}^*_{-z_G} + B_G = \hat{f} + A_G + B_G.$$

□

13.1.1 数据添加

定理 13.3。对于数据集 $\mathcal{D} = \{(x_i,\ y_i,\ c_i)\}_{i=1}^N$，给定一组待添加的新数据 $\mathcal{D}_{\text{add}} = \{\tilde{z}_s = (\tilde{x}_s,\ \tilde{y}_s,\ \tilde{c}_s)\}_{s=1}^M$。假设更新后的概念预测器 $\hat{g}^{\cup\mathcal{D}_{add}}$ 定义为

$$\hat{g}^{\cup\mathcal{D}_{add}} = \arg\min_g \sum_{j\in[K]} \sum_{i\in[N]} L_{C_j}(g(x_i), c_i) + \sum_{j\in[K]} \sum_{s\in[M]} L_{C_j}(g(\tilde{x}_S), \tilde{c}_S). \quad (83)$$

定义 $L_C(\hat{}g(x_i)\ c_i) \triangleq \sum_{j=1}^{K} L_{C_j}(\hat{}g(x_i)\ c_i)$。那么我们可以得到 $\hat{}g^{\cup\mathcal{D}_{add}}$ 的如下近似

$$\hat{g}^{\cup\mathcal{D}_{add}} \approx \bar{g}^{\cup\mathcal{D}_{add}} \triangleq \hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{s\in[M]} \nabla_g L_C(\hat{g}(\tilde{x}_s), \tilde{c}_s), \quad (84)$$

其中 $H_{\hat{g}} = \nabla_{\hat{g}}^2 \sum_{i=1}^{N} L_C(\hat{}g(x_i)\ c_i)$ 是损失函数关于 $\hat{}g$的海森矩阵。

证明。 首先，回顾 $\hat{}g^{\cup\mathcal{D}_{\text{add}}}$ 的定义：

$$\hat{g}^{\cup\mathcal{D}_{\text{add}}} = \arg\min_g \left[\sum_{i=1}^{N} L_C(g(x_i), c_i) + \sum_{s=1}^{M} L_C(g(\tilde{x}_s), \tilde{c}_s)\right],$$

然后，我们将第 s 个数据的权重增加 ϵ，得到一个新的预测器 $\hat{}g_\epsilon^{\cup\mathcal{D}_{\text{add}}}$:

$$\arg\min_g \left[\sum_{i=1}^{N} L_C(g(x_i), c_i) + \epsilon \cdot \sum_{s=1}^{M} L_C(g(\tilde{x}_s), \tilde{c}_s)\right]. \quad (85)$$

因为 $\hat{}g_\epsilon^{\cup\mathcal{D}_{\text{add}}}$使方程（85）右侧最小，因此我们有

$$\nabla_g \sum_{i=1}^{N} L_Y(\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}}(x_i), c_i) + \epsilon \cdot \nabla_g \sum_{s=1}^{M} L_Y(\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}}(\tilde{x}_s), \tilde{c}_s) = 0.$$

当 $\epsilon \to 0$ 时，$\hat{}g_\epsilon^{\cup\mathcal{D}_{\text{add}}} \to \hat{}g$。因此，我们可以关于 g 进行一阶泰勒展开，得到

$$\begin{aligned} 0 \approx & \nabla_g \sum_{i=1}^{N} L_C(\hat{g}(x_i), c_i) + \epsilon \cdot \nabla_g \sum_{s=1}^{M} L_C(\hat{g}(\tilde{x}_s), \tilde{c}_s) \\ & + \nabla_g^2 \sum_{i=1}^{N} L_C(\hat{g}(x_i), c_i) \cdot (\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}} - \hat{g}). \end{aligned} \quad (86)$$

回顾 $\hat{}g$ 的定义：

$$\hat{g} = \arg\min_g \sum_{i=1}^{n} L_Y(g(x_i), c_i),$$

那么，(86) 右侧的第一项等于 0。令$\epsilon \to 0$，则有

$$\left.\frac{\mathrm{d}\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}}}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = -H_{\hat{g}}^{-1} \cdot \sum_{s=1}^{M} \nabla_g L_C(\hat{g}(\tilde{x}_s), \tilde{c}_s),$$

其中 $H_{\hat{g}} = \nabla_{\hat{g}}^2 \sum_{i=1}^{N} L_C(\hat{}g(x_i)\ c_i)$。

注意，当 $\epsilon \to 0$ 时，$\hat{}g_\epsilon^{\cup\mathcal{D}_{\text{add}}} \to \hat{}g^{\cup\mathcal{D}_{\text{add}}}$。在 $\hat{}g$ 处执行一个牛顿步，便得到在概念层级上编辑原始概念预测器的方法：

$$\hat{g}^{\cup\mathcal{D}_{\text{add}}} \approx \bar{g}^{\cup\mathcal{D}_{\text{add}}} \triangleq \hat{g} - H_{\hat{g}}^{-1} \cdot \sum_{s=1}^{M} \nabla_g L_C(\hat{g}(\tilde{x}_s), \tilde{c}_s).$$

□

定理13.4。 对于数据集 $\mathcal{D} = (x_i,\ y_i,\ c_i)i = 1^N$，给定一组待添加的新数据 $\mathcal{D}$添加 $= \tilde{z}s = (\tilde{x}s,\ \tilde{y}s,\ \tilde{c}s)s = 1^M$。假设更新后的标签预测器 $\hat{}f^{\cup\mathcal{D}_{add}}$ 定义如下

$$\begin{aligned} \hat{f}^{\cup\mathcal{D}_{add}} = & \arg\min_f \sum_{i=1}^{N} L_Y(f(\hat{g}^{\cup\mathcal{D}_{add}}(x_i)), y_i) \\ & + \sum_{s=1}^{M} L_Y(f(\hat{g}^{\cup\mathcal{D}_{add}}(\tilde{x}_s)), \tilde{y}_s). \end{aligned} \quad (87)$$

那么，我们可以得到 $\hat{}f^{\cup\mathcal{D}_{add}}$的如下近似：

$$\hat{f}^{\cup\mathcal{D}_{add}} \approx \bar{f}^{\cup\mathcal{D}_{add}} \triangleq \hat{f} - H^{-1}\hat{f} \cdot \sum_{s\in[M]} \nabla_f L_Y(\hat{f}(\tilde{x}s), \tilde{y}s), \quad (88)$$

其中 $H\hat{}f = \nabla^2\hat{}f\sum_{i=1}^{N} L_Y(\hat{}f(x_i)\ y_i)$ 是损失函数关于 f 的海森矩阵。

证明。 我们可以看到，$\hat{}f^{\cup\mathcal{D}_{\text{add}}}$与$\hat{}f$之间存在巨大差距。因此，首先，我们定义 $\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$ 为

$$\arg\min_f \sum_{i=1}^{N} L_Y\left(f(\hat{g}(x_i)), y_i\right) + \sum_{s=1}^{M} L_Y\left(f(\hat{g}(\tilde{x}_s)), \tilde{y}_s\right).$$

然后，我们将 $\tilde{f}_\epsilon^{\cup\mathcal{D}_{\text{add}}}$定义如下，以估计 $\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$。

$$\begin{aligned} \tilde{f}_\epsilon^{\cup\mathcal{D}_{\text{add}}} = & \arg\min_f \sum_{i=1}^{N} L_Y\left(f(\hat{g}(x_i)), y_i\right) \\ & + \epsilon \cdot \sum_{s=1}^{M} L_Y\left(f(\hat{g}(\tilde{x}_s)), \tilde{y}_s\right). \end{aligned}$$

根据最小化条件，我们有

$$\begin{aligned} 0 = & \nabla_f \sum_{i=1}^{N} L_Y\left(\tilde{f}_\epsilon^{\cup\mathcal{D}_{\text{add}}}(\hat{g}(x_i)), y_i\right) \\ & + \epsilon \cdot \sum_{s=1}^{M} \nabla_f L_Y\left(\tilde{f}_\epsilon^{\cup\mathcal{D}_{\text{add}}}(\hat{g}(\tilde{x}_s)), \tilde{y}_s\right). \end{aligned}$$

$\hat{}$在 f处进行一阶泰勒展开，

$$\begin{aligned} & \nabla_{\hat{f}} \sum_{i=1}^{N} L_Y\left(\hat{f}(\hat{g}(x_i)), y_i\right) + \epsilon \cdot \nabla_{\hat{f}} \sum_{s=1}^{M} L_Y\left(\hat{f}(\hat{g}(\tilde{x}_s)), \tilde{y}_s\right) \\ & + \nabla_{\hat{f}}^2 \sum_{i=1}^{N} L_Y\left(\hat{f}(\hat{g}(x_i)), y_i\right)\left(\tilde{f}_\epsilon^{\cup\mathcal{D}_{\text{add}}} - \hat{f}\right) = 0. \end{aligned}$$

然后， $\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$ 可以近似为

$$\tilde{f}^{\cup\mathcal{D}_{\text{add}}} \approx \hat{f} - H_{\hat{f}}^{-1} \cdot \sum_{s=1}^{M} \nabla_{\hat{f}} L_Y\left(\hat{f}(\hat{g}(\tilde{x}_s)), \tilde{y}_s\right). \quad (89)$$

定义 $A \triangleq H^{-1}\hat{}_f \cdot \sum_{s=1}^{M} \nabla\hat{}_f L_Y\left(\hat{}f(\hat{}g(\tilde{x}_s))\ \tilde{y}_s\right)$然后，$\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$的近似定义为

$$\bar{f}^{\cup\mathcal{D}_{\text{add}}} = \hat{f} - A \quad (90)$$

然后，我们估计 $\hat{}f^{\cup\mathcal{D}_{\text{add}}}$ 与 $\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$ 之间的差异。回顾 $\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$ 的定义：

$$\arg\min_f \sum_{i=1}^{N} L_Y\left(f(\hat{g}(x_i)), y_i\right) + \sum_{s=1}^{M} L_Y\left(f(\hat{g}(\tilde{x}_s)), \tilde{y}_s\right). \quad (91)$$

比较方程（87）和（91），我们仍需将中间预测器 $\hat{}f_{s_t}^{\cup\mathcal{D}_{\text{add}}}$定义为

$$\begin{aligned} \hat{f}_{s_t}^{\cup\mathcal{D}_{\text{add}}} = \arg\min_f \Bigg[& \sum_{i=1}^{N} L_Y\left(f(\hat{g}(x_i)), y_i\right) \\ & + \sum_{s=1}^{M} L_Y\left(f(\hat{g}(\tilde{x}_s)), \tilde{y}_s\right) \\ & + L_Y\left(f(\hat{g}_\epsilon^{\cup\mathcal{D}_{\text{add}}}(\tilde{x}_{s_t})), \tilde{y}_{s_t}\right) \\ & - L_Y\left(f(\hat{g}(\tilde{x}_{s_t})), \tilde{y}_{s_t}\right)\Bigg] \end{aligned}$$

用某个 ϵ 对 s_t 数据进行加权提升，我们将 $\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}}$ 定义为

$$\begin{aligned}\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}} = \arg\min_{f}\Bigg[&\sum_{i=1}^{N} L_Y\left(f(\hat{g}(x_i)), y_i\right)\\&+\sum_{s=1}^{M} L_Y\left(f(\hat{g}(\tilde{x}_s)), \tilde{y}_s\right)\\&+\epsilon\cdot L_Y\left(f(\hat{g}^{\cup\mathcal{D}_{\text{add}}}(\tilde{x}_{s_t})), \tilde{y}_{s_t}\right)\\&-\epsilon\cdot L_Y\left(f(\hat{g}(\tilde{x}_{s_t})), \tilde{y}_{s_t}\right)\Bigg].\end{aligned}$$

因此，根据最小化条件，我们有

$$\begin{aligned}0 = \nabla_f \sum_{i=1}^{N} L_{Y_i}\left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right) + \sum_{s=1}^{M} L_{Y_i}\left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\\ +\epsilon\cdot\nabla_f L_{Y_{st}}\left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}^{\cup\mathcal{D}_{\text{add}}}\right) - \epsilon\cdot\nabla_{\hat{f}} L_{Y_{ir}}\left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right).\end{aligned} \tag{92}$$

当 $\epsilon\to 0$，$\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}} \to \tilde{f}^{\cup\mathcal{D}_{\text{add}}}$。然后，我们在 $\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$ 处对方程 92进行泰勒展开，得到

$$\begin{aligned}0 \approx \nabla_f \sum_{i=1}^{N} L_{Y_i}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right) + \sum_{s=1}^{M} L_{Y_i}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\\ +\epsilon\cdot\nabla_f L_{Y_{st}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}^{\cup\mathcal{D}_{\text{add}}}\right)\cdot\left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}}\right)\\ -\epsilon\cdot\nabla_{\hat{f}} L_{Y_{ir}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\cdot\left(\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}}\right).\end{aligned}$$

整理上述方程可得

$$\begin{aligned}\hat{f}_{s_t,\epsilon}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}} \approx -\epsilon\cdot H_{\tilde{f}}^{-1}\nabla_f\Big(L_{Y_{st}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}^{\cup\mathcal{D}_{\text{add}}}\right)\\ -L_{Y_{st}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\Big),\end{aligned}$$

其中

$$H_{\tilde{f}} = \nabla_{\hat{f}}^2 \sum_{i=1}^{N} L_{Y_i}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right) + \sum_{s=1}^{M} L_{Y_i}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right).$$

当 $\epsilon = 1$ 时，$\hat{f}_{s_t}^{\cup\mathcal{D}_{\text{add}},\,-1} = \hat{f}_{s_t}^{\cup\mathcal{D}_{\text{add}}}$。然后，我们在 $\tilde{f}^{\cup\mathcal{D}_{\text{add}}}$ 处以步长 1进行牛顿迭代，

$$\begin{aligned}\hat{f}_{s_t}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}} \approx -H_{\tilde{f}}^{-1}\nabla_f\Big(L_{Y_{st}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}^{\cup\mathcal{D}_{\text{add}}}\right)\\ -L_{Y_{st}}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\Big),\end{aligned}$$

遍历集合 $[M]$ 中的 s_t，我们得到

$$\begin{aligned}\hat{f}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}} \approx -H_{\tilde{f}}^{-1}\cdot\sum_{s=1}^{M}\nabla_f\Bigg(L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}^{\cup\mathcal{D}_{\text{add}}}\right)\\ - L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\Bigg)\end{aligned} \tag{93}$$

编辑后的 $\hat{g}^{\cup\mathcal{D}_{\text{add}}}$ 已根据 $\bar{g}^{\cup\mathcal{D}_{\text{add}}}$定理13.3推导得出，将此近似代入（93），可得

$$\begin{aligned}\hat{f}^{\cup\mathcal{D}_{\text{add}}} - \tilde{f}^{\cup\mathcal{D}_{\text{add}}} \approx -H_{\tilde{f}}^{-1}\cdot\nabla_{\hat{f}}\sum_{s=1}^{M}\Big[L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \bar{g}^{\cup\mathcal{D}_{\text{add}}}\right)\\ - L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\Big].\end{aligned} \tag{94}$$

注意，由于我们没有重新训练标签预测器，而是将其编辑为 $\bar{f}^{\cup\mathcal{D}_{\text{add}}}$ 作为替代，因此无法直接得到 $\hat{f}^{\cup\mathcal{D}_{\text{add}}}$ 和 $H_{\tilde{f}}$。因此，我们用 $\bar{f}^{\cup\mathcal{D}_{\text{add}}}$ 近似 $\hat{f}^{\cup\mathcal{D}_{\text{add}}}$，并用 $H_{\bar{f}}$ 近似 $H_{\tilde{f}}$，其定义如下：

$$H_{\bar{f}} = \nabla_{\bar{f}}^2 \sum_{i=1}^{N} L_{Y_i}\left(\bar{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right) + \nabla_{\bar{f}}^2 \sum_{s=1}^{M} L_{Y_s}\left(\bar{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)$$

然后我们将 B 定义为

$$B \triangleq H_{\bar{f}}^{-1}\cdot\nabla_{\hat{f}}\left(\sum_{s=1}^{M} L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \bar{g}^{\cup\mathcal{D}_{\text{add}}}\right) - L_{Y_s}\left(\tilde{f}^{\cup\mathcal{D}_{\text{add}}}, \hat{g}\right)\right) \tag{95}$$

合并（90）和（82），然后我们推导出最终的闭式编辑后的标签预测器，如下所示

$$\bar{f}_*^{\cup\mathcal{D}_{\text{add}}} = \bar{f}^{\cup\mathcal{D}_{\text{add}}} - B = \hat{f} - A - B.$$

□

14 现有方法

本节介绍用于加速影响力计算的 EK-FAC 方法。

14.1 EK-FAC

在我们的概念瓶颈模型中，标签预测器是一个单个线性层，计算海森矩阵的成本可以承受。然而，概念预测器基于 ResNet-18，参数多得多。因此，我们对 $\hat{}g$ 使用 EK-FAC。

$$\hat{g} = \arg\min_{g} \sum_{j=1}^{k} L_{C_j} = \arg\min_{g} \sum_{j=1}^{k} \sum_{i=1}^{n} \ell(g^j(x_i), c_i^j),$$

我们将 $H_{\hat{g}} = \nabla_{\hat{g}}^2 \sum L_{C_j}$ 定义为损失函数关于参数的海森矩阵。

为此，考虑 $\hat{}g$ 的第 l 层，该层以激活值层 $\{a_{j,t}\}$ 作为输入，其中 $j \in \{1, 2, \ldots, J\}$ 表示输入映射的索引，$t \in \mathcal{T}$ 表示空间位置的索引，空间位置通常构成二维网格。该层由一组权重 $W = (w_{i,\ j,\ \delta})$ 和偏置 $b = (b_i)$ 参数化，其中 $i \in \{1, \ldots, I\}$ 表示输出映射的索引，$\delta \in \Delta$ 表示空间偏移的索引（相对于滤波器中心）。

卷积层计算一组预激活值，表示为

$$[S_l]_{i,t} = s_{i,t} = \sum_{\delta \in \Delta} w_{i,j,\delta} a_{j,t+\delta} + b_i.$$

将相对于 $s_{i,\ t}$ 的损失导数记为

$$\mathcal{D}s_{i,t} = \frac{\partial \sum L_{C_j}}{\partial s_{i,t}},$$

它可以在反向传播过程中计算得到。

激活值实际存储为维度为 $|\mathcal{T}| \times J$ 的 A_{l-1}。同样，权重存储为一个 $I \times |\Delta| J$ 数组 W_l。尽管理论上卷积的直接实现具有高度并行性，但其内存访问模式较差。因此，高效实现通常会利用一种称为扩展算子 $[\![\cdot]\!]$ 的方法。例如，$[\![A_{l-1}]\!]$ 是一个 $|\mathcal{T}| \times J|\Delta|$ 矩阵，定义如下：

$$[\![A_{l-1}]\!]_{t,j|\Delta|+\delta} = [A_{l-1}]_{(t+\delta),j} = a_{j,t+\delta},$$

为了将偏置合并到权重中，我们需要在扩展后的激活值 $[\![A_{l-1}]\!]$ 中添加一个齐次坐标（即一列全为1的值），并将其记为 $[\![A_{l-1}]\!]_{\mathrm{H}}$。将偏置向量与权重矩阵拼接后，我们得到 $\theta_l = (b_l,\ W_l)$。

那么，$H_{\hat{g}}$ 的近似为：

$$\begin{aligned} G^{(l)}(\hat{g}) =& \mathbb{E}\left[\mathcal{D}w_{i,j,\delta} \mathcal{D}w_{i',j',\delta'}\right] \\ =& \mathbb{E}\left[\left(\sum_{t \in \mathcal{T}} a_{j,t+\delta} \mathcal{D}s_{i,t}\right)\left(\sum_{t' \in \mathcal{T}} a_{j',t'+\delta'} \mathcal{D}s_{i',t'}\right)\right] \\ \approx& \mathbb{E}\left[[\![A_{l-1}]\!]_{\mathrm{H}}^{\top} [\![A_{l-1}]\!]_{\mathrm{H}}\right] \\ & \otimes \frac{1}{|\mathcal{T}|} \mathbb{E}\left[\mathcal{D}S_l^{\top} \mathcal{D}S_l\right] \triangleq \Omega_{l-1} \otimes \Gamma_l. \end{aligned}$$

使用训练集的均值估计期望，

$$\begin{aligned} G^{(l)}(\hat{g}) \approx& \frac{1}{n} \sum_{i=1}^{n} \left([\![A_{l-1}^i]\!]_{\mathrm{H}}^{\top} [\![A_{l-1}^i]\!]_{\mathrm{H}}\right) \\ & \otimes \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{|\mathcal{T}|} \mathcal{D}S_l^{i\top} \mathcal{D}S_l^i\right) \triangleq \hat{\Omega}_{l-1} \otimes \hat{\Gamma}_l \end{aligned}$$

此外，如果因子 $\hat{}\Omega_{l-1}$ 和 $\hat{}\Gamma_l$ 的特征分解分别为 $Q_\Omega \Lambda_\Omega Q_\Omega^\top$ 和 $Q_\Gamma \Lambda_\Gamma Q_\Gamma^\top$，那么 $\hat{}\Omega_{l-1} \otimes \hat{}\Gamma_l$ 的特征分解可写为：

$$\begin{aligned} \hat{\Omega}_{l-1} \otimes \hat{\Gamma}_l &= Q_\Omega \Lambda_\Omega Q_\Omega^\top \otimes Q_\Gamma \Lambda_\Gamma Q_\Gamma^\top \\ &= (Q_\Omega \otimes Q_\Gamma)(\Lambda_\Omega \otimes \Lambda_\Gamma)(Q_\Omega \otimes Q_\Gamma)^\top \end{aligned}$$

由于后续需要进行求逆运算，且当前对 $G^{(l)}(\hat{}g)$ 的近似是半正定的，因此我们实际使用阻尼版本，形式如下：

$$\hat{G}^l(\hat{g})^{-1} = (G_l(\hat{g}) + \lambda_l I_{d_l})^{-1} \tag{96}$$

$$= (Q_{\Omega_{l-1}} \otimes Q_{\Gamma_l})(\Lambda_{\Omega_{l-1}} \otimes \Lambda_{\Gamma_l} + \lambda_l I_{d_l})^{-1} (Q_{\Omega_{l-1}} \otimes Q_{\Gamma_l})^{\mathrm{T}} \tag{97}$$

此外，[37] 提出了一种新方法，用于修正方程96 中的错误。该方法将 i-th 对角元素 $\Lambda_{\Omega_{l-1}} \otimes \Lambda_{\Gamma_l}$ 设为

$$\Lambda_{ii}^* = n^{-1} \sum_{j=1}^{n} \left((Q_{\Omega_{l-1}} \otimes Q_{\Gamma_l}) \nabla_{\theta_l} \ell_j\right)_i^2.$$

15 更多可视化结果与解释

可视化。由于 CBM 是一种可解释模型，我们旨在评估 CCBM 的可解释性（与重新训练方法相比）。我们将展示概念层级编辑的一些可视化结果。图10 展示了在 CUB数据集上随机删除概念后，由我们的 CCBM 和重新训练方法得到的排名前 10 个最具影响力的概念及其对应的预测概念标签。（详细解释见附录15.1。）我们的 CCBM 能够解释哪些概念至关重要，以及这些概念如何辅助预测。具体而言，在真实值（重新训练方法）中最重要的前 10 个概念里，CCBM 能够准确识别出其中 9 个。例如，在预测类别时，我们正确识别出“has upperparts color::orange”“has upper tail color::red”和“has breast color::black”是最重要的概念之一。在附录15.2中还提供了 OAI 和 CUB数据集上数据级别和概念标签层级的其他可视化结果。



图10：在提取的图像上突出显示概念瓶颈模型（由 CCBM 或重新训练识别）的影响力最大的前10个概念的可视化结果。

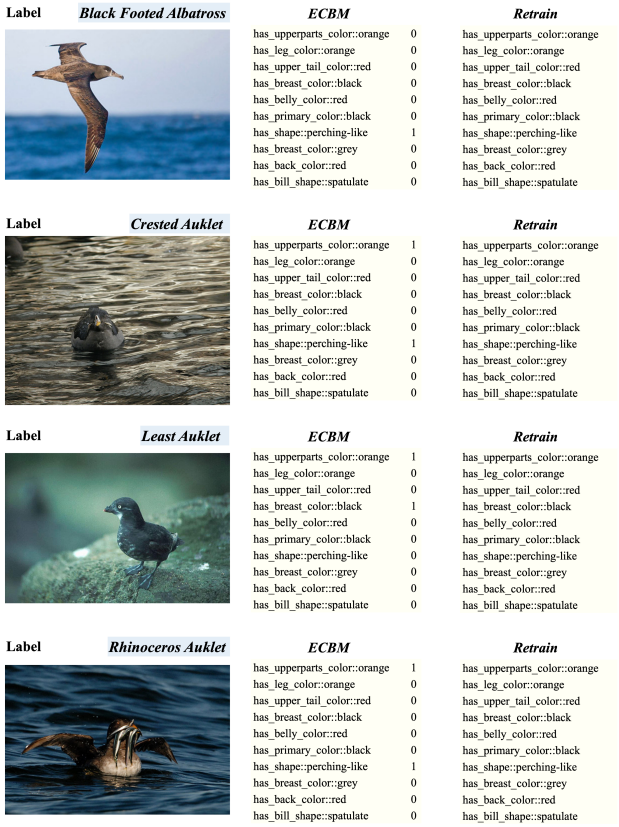


图 11：CUB 中不同类别排名前 10 的最重要概念的可视化结果。

15.1 可视化结果说明

在概念层级，我们每次移除一个概念，重新训练概念瓶颈模型，随后评估模型性能。我们根据模型性能损失从高到低对概念进行排序。移除后会导致模型性能显著变化的概念被视为有影响力的概念。如图 10 所示，重新训练列展示了排名前10的概念。相比之下，我们采用自己的 CCBM 方法，而非重新训练方法，如算法 8 所述；图 10 的 CCBM 列展示了排名前10的概念。

为帮助读者将排名前 10 的重要概念与输入图像联系起来，我们提供了数据的可视化结果，并列出了与排名前 10 的重要概念对应的概念标签，如图 10、11、12所示。

对于另外两个层级以及其他数据集，我们也执行了类似的流程，相应的可视化结果如图13、14、15、16和17所示。

15.2 可视化结果

我们在图 11、12、13、14、15、16和17中展示了额外的可视化结果。

16 更多相关工作

影响函数。影响函数最初是稳健统计学中的常用工具，[22], [23], 自 [38] 将其引入该领域以来，便在机器学习中得到了广泛应用。其多功能性涵盖了各种应用，包括检测错误标注数据、模型解释、处理模型偏差，以及促进机器遗忘任务。机器遗忘领域的代表性工作包括特征与标签遗忘 [39], 极小极大遗忘 [40], 遗忘用于训练深度神经网络的部分图像数据 [41], [42],以及涉及节点、边和特征的图遗忘。近期的进展，例如 LiSSA 方法 [43], [44]和基于 kNN 的技术 [45], ，已被提出以提升计算效率。此外，各种研究已应用影响函数来解释不同领域中的模型，包括自然语言处理[46]和图像分类 [47],，同时还处理分类模型中的偏置 [48], 词嵌入 [49],以及微调模型 [50]。尽管已有大量关于影响函数的研究，我们还是首次利用它们构建可编辑的概念瓶颈模型。与传统神经网络相比，概念瓶颈模型中的影响函数更为复杂。这是因为在传统影响函数中，我们只需要改变预测输出。而在概念瓶颈模型中，我们首先应移除真实概念，然后需要近似预测概念，以便近似输出。在真实概念与预测概念之间弥合差距，对我们的证明构成了重大的理论挑战。

模型遗忘。近年来，模型遗忘受到了广泛关注，研究者提出了多种方法 [51]–[55]，以高效移除训练后的机器学习模型中某些数据的影响。现有方法大致可分为精确遗忘和近似遗忘



图12：CUB中不同类别最具影响力的前10个概念的可视化。

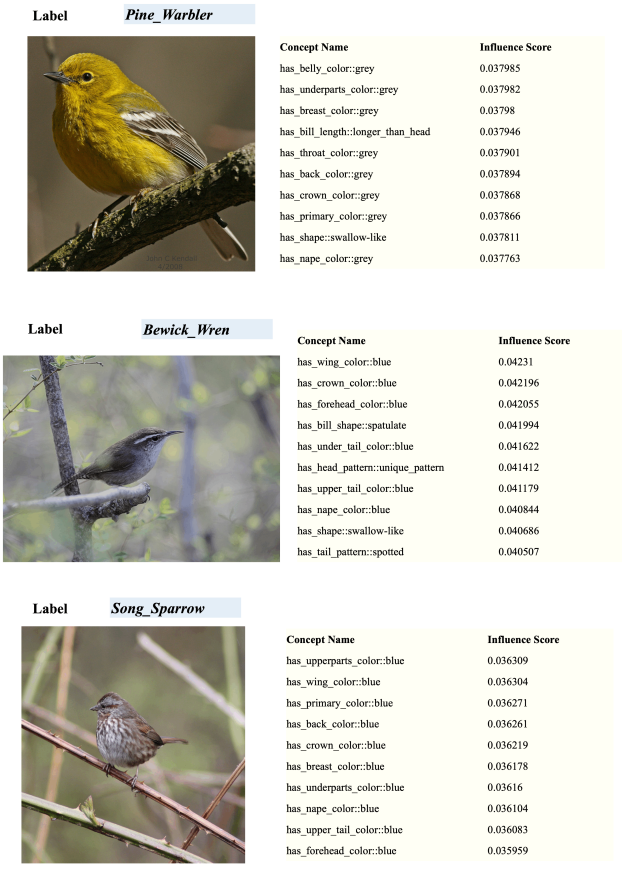


图13：CUB中与不同数据相关的最具影响力概念标签的可视化。

方法。精确遗忘方法旨在通过有选择地更新数据集的一部分来复现重新训练的结果，从而避免在整个数据集上重新训练所带来的计算开销 [56], [57]。另一方面，近似遗忘方法则试图调整模型参数，使其在剩余数据上近似满足目标函数的最优性条件 [41], [58], [59]。这些方法进一步分为三个子类别：（1）利用海森矩阵相关项的基于牛顿步的更新 [22, 26, 31, 34, 40, 43, 49], ，通常会加入高斯噪声以减轻残余数据影响。为降低计算成本，一些研究使用费 Fisher 信息矩阵 [41] 或较小的海森矩阵块 [60]来近似海森矩阵。（2）基于神经切线核（NTK）的机器遗忘将训练近似为线性过程，可以将其视为单次线性变化 [61]。（3）DeltaGrad [62]和 unrollSGD [63],等 SGD 路径跟踪方法会逆转随机梯度下降在训练期间的优化轨迹。尽管这些方法取得了进展，但它们无法处理概念瓶颈模型的特殊架构。此外，考虑到获取数据的成本很高，有时我们更希望纠正数据，而不是将其移除；而模型遗忘无法实现这一点。

17 局限性与更广泛的影响

17.1 局限性

尽管 CCBM 展现出显著的效率和实用性，但也必须承认该方法固有的一些局限性

。首先，**非凸情形下的近似误差**：我们的方法依赖于影响函数，而影响函数基于损失函数的二阶泰勒展开。这一方法假设损失景观在局部是凸的、呈二次曲线。在损失曲面高度非凸的深度神经网络中，这种近似可能偏离精确的重新训练解，尤其是在编辑幅度较大时（例如，删除或添加的样本数量较多）。尽管我们的序贯编辑实验表明该方法具有鲁棒性，但极其激进的修改仍可能需要定期进行完全重新训练，以消除累积误差。其次，**海森矩阵的计算开销**：尽管使用 EK-FAC 加速计算，计算和求逆海森矩阵（或费 Fisher 信息矩阵）仍会产生内存和计算开销，且开销会随参数的维度而增加。对于参数量达数十亿的超大规模基础模型，可能需要进一步优化或采用稀疏近似，才能使 CCBMs 完全实现实时运行。

17.2 更广泛的影响

我们的工作不仅仅是纠正模型错误，还为可解释人工智能系统的**生命周期管理**提供了基础框架。

通过数据添加实现持续适应。静态模型不适用于动态环境。CCBM 的一项关键

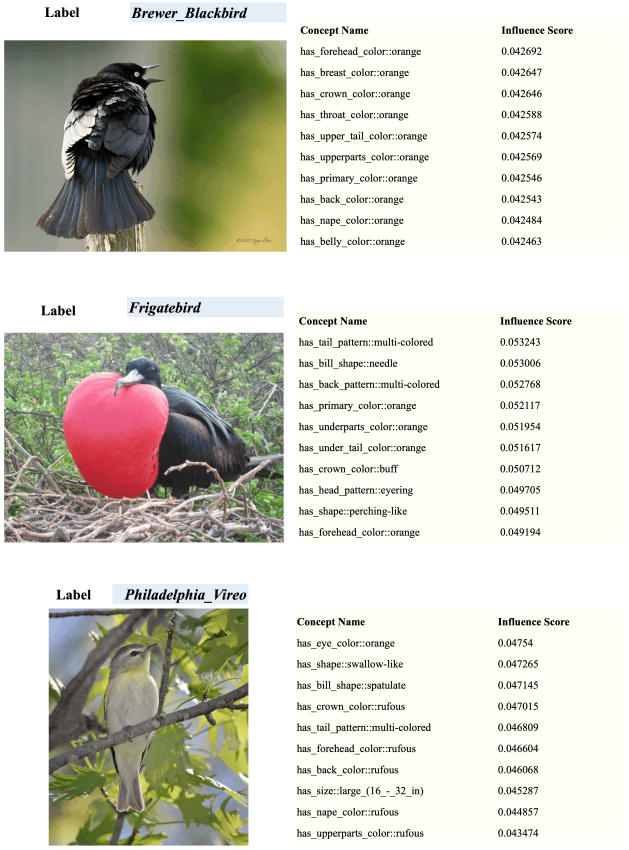


图14：与CUB中不同数据相关的最具影响力概念标签的可视化。

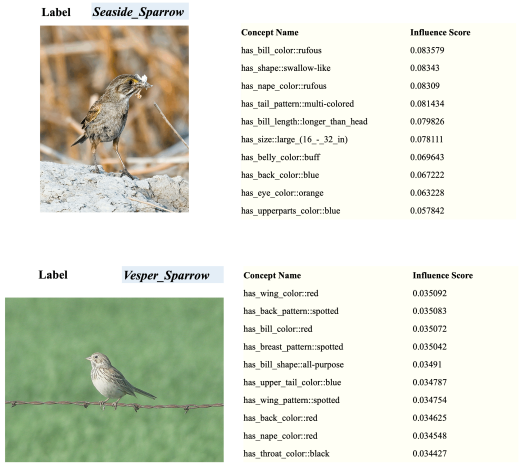


图15：与CUB中不同数据相关的最具影响力概念标签的可视化。

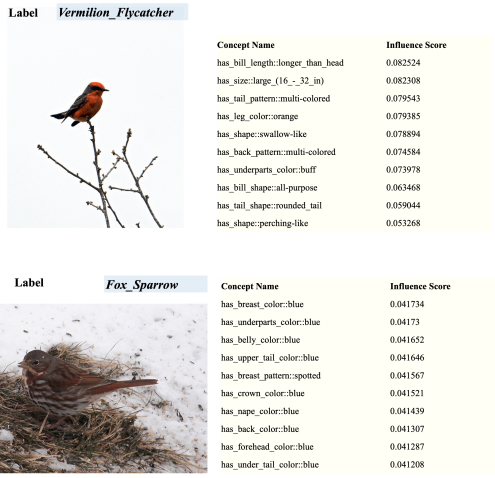


图16：与CUB中不同数据相关的最具影响力概念标签的可视化。

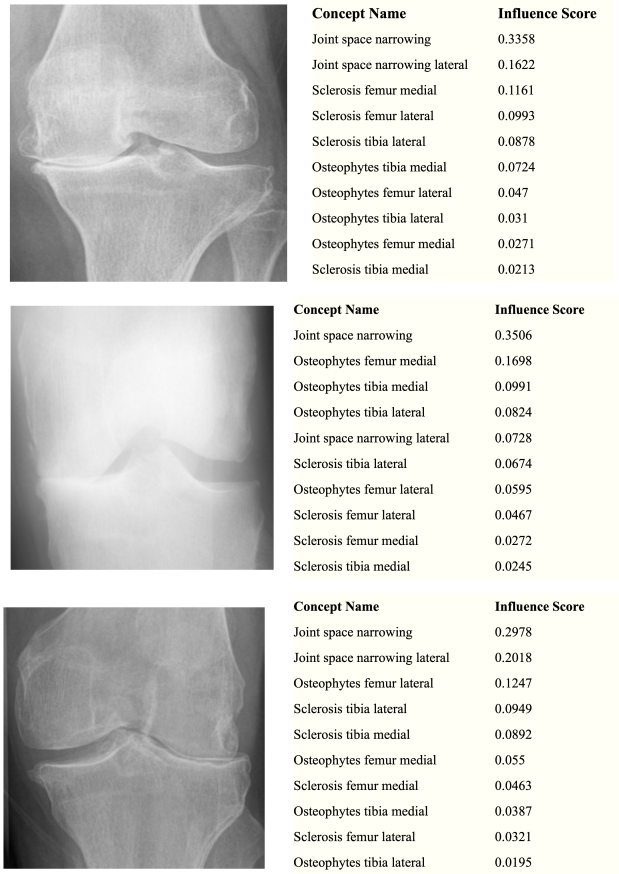


图17：与OAI中不同数据相关的最具影响力概念标签的可视化。

贡献在于支持**数据添加**，使模型能够高效地进行增量学习。在医疗保健或金融等领域，数据会以流的形式不断涌入。CCBMs 使已部署的模型能够持续吸收新案例（例如，疾病的新变体或新型金融欺诈模式），而无需承受重新训练带来的停机和能耗。这改变了概念瓶颈模型从静态产物发展为与人类知识共同成长的演化系统。

助力隐私保护与合规。 随着《通用数据保护条例》（GDPR）和《加州消费者隐私法案》（CCPA）等法规的实施，“被遗忘权”已成为一项法律要求。CCBM 提供了一种有坚实数学依据的机制

用于**数据删除**，使组织能够严格地对敏感用户数据进行机器遗忘，或清除中毒样本。我们的 RMIA 实验验证了这种机器遗忘在降低隐私风险方面的有效性，从而增强人们对人工智能部署的信任。**推动可持续的“绿色人工智能”**。 重新训练现代深度学习模型会消耗大量资源，并产生显著的碳足迹。通过以高效的闭

-

更新形式后，CCBMs 大幅降低了能源消耗 -

（例如，如我们的实验所示，将运行时间从数小时缩短至数分钟）。这符合绿色人工智能的目标，使高性能可解释模型在环境方面更加可持续。

赋能人机协作。最后，CCBMs 充当了实现交互式人机对齐的桥梁。通过允许专家（例如医生）在概念标签或概念层级进行干预，模型让用户能够直接纠正推理缺陷。这种“可编辑”的特性确保模型始终与专家共识保持一致，显著提升其在关键决策流程中的可靠性和应用度。