条件熵正则项（CEL）在 CEM-main 与 gated-att 中的计算机制详解

2025 年 10 月 30 日

目录

1 引言与整体目标

本文面向 `CEM-main` 与 `gated-att` 两个项目，剖析其在协同推理（split inference）场景下用于防御模型反演攻击的条件熵正则项（Conditional Entropy Loss, CEL）的全流程计算方式。为方便比较与复现，我们从“输入数据形态、统计量构造、损失公式、梯度回传策略”四个角度细致梳理两种方法的异同，并给出一个可手算的小规模示例用于说明。

在整个系统中，客户端编码器 $f_{\boldsymbol{\theta}}$ 将输入图像 x 映射为被称为 smashed data 的中间表示 $z = f_{\boldsymbol{\theta}}(x)$。服务器端根据 z 接续推理，但攻击者也可能利用 z 重建出原始图像，构成隐私威胁。CEL 的设计目标是：**约束同类 smashed data 在特征空间中尽量紧凑**，降低攻击者从 z 重建出原始输入的可行性。

2 统一符号约定

- 输入样本 (x, y)，其中 $x \in \mathbb{R}^{h\times w\times c}$，标签 $y \in \{1, 2, \ldots, C\}$。
- 客户端编码器 $f_{\boldsymbol{\theta}} : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^d$，输出 $z = f_{\boldsymbol{\theta}}(x)$。
- 当前 batch 的 smashed data 记为 $Z_B = \{z_i\}_{i=1}^{|B|}$，标签为 $Y_B = \{y_i\}_{i=1}^{|B|}$。
- 全局训练集中类别 c 的样本集合记为 $\mathcal{D}^{(c)}$，在某轮训练得到的 smashed data 全量集合为 Z_{all}。
- 分类损失为 $\mathcal{L}_{\text{CE}}$，CEL 记为 $\mathcal{L}_{\text{CEL}}$，总损失 $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{CEL}}$。
- 超参数：簇数 K、log-entropy 阈值 τ、噪声方差上界 σ^2、warm-up 轮数 T_{warm} 等。

3 `CEM-main` 中的 CEL 计算流程

3.1 总体思路

`CEM-main` 以“**离线聚类 + 方差门控**”为核心。即在每个 epoch 结束后，收集所有 smashed data，按类别执行 K-means 或混合高斯（GMM）聚类，借助簇方差衡量 smashed data 的分布紧致度。下一轮训练时，批内的 smashed data 通过聚类结果获得其类内方差估计，从而形成 CEL。

3.2 原始数据与统计量构建

步骤 1：全量 smashed data 缓存。 设当前 epoch 中遍历完所有 mini-batch 后得到的 smashed data 集合为

$$Z_{\text{all}} = \{(z_i, y_i)\}_{i=1}^{N}, \quad z_i = f_{\boldsymbol{\theta}}(x_i),$$

其中 N 为当轮训练样本总数。数据被分门别类地组织为

$$Z_{\text{all}}^{(c)} = \{z_i \mid y_i = c\}.$$

步骤 2：按类聚类。 对每个类别 c，执行带暖启动的 K-means：

初始化簇中心$\{\mu_{c,k}^{(0)}\}_{k=1}^{K}$（可使用上轮结果）。

repeat

分配：$S_{c,k}^{(t)} = \left\{z \in Z_{\text{all}}^{(c)} \mid k = \arg\min_{k'} \|z - \mu_{c,k'}^{(t)}\|_2^2\right\}$

更新：$\mu_{c,k}^{(t+1)} = \dfrac{1}{|S_{c,k}^{(t)}|} \displaystyle\sum_{z \in S_{c,k}^{(t)}} z$

until 收敛或达到最大迭代次数

得到簇中心 $\mu_{c,k}$，进一步估计每个簇的方差与协方差：

$$v_{c,k} = \frac{1}{|S_{c,k}|} \sum_{z \in S_{c,k}} \|z - \mu_{c,k}\|_2^2, \tag{1}$$

$$\Sigma_{c,k} \approx \operatorname{diag}\left(\frac{1}{|S_{c,k}|} \sum_{z \in S_{c,k}} (z - \mu_{c,k})(z - \mu_{c,k})^{\top}\right). \tag{2}$$

同时记录簇权重 $\pi_{c,k} = \frac{|S_{c,k}|}{|Z_{\text{all}}^{(c)}|}$。

若采用 GMM，则借助 EM 算法更新高斯混合参数 $(\pi_{c,k}, \mu_{c,k}, \Sigma_{c,k})$。无论 K-means 还是 GMM，最后都得到一组可复用的统计量：

$$\mathcal{S}_c = \{(\pi_{c,k}, \mu_{c,k}, \Sigma_{c,k})\}_{k=1}^{K}.$$

3.3 训练迭代中的 CEL 估计

在下一轮训练的任意一个 mini-batch 中，考虑类别 c 的子集

$$Z_B^{(c)} = \{z_i \in Z_B \mid y_i = c\}.$$

步骤 1：簇匹配。 对每个 $z \in Z_B^{(c)}$，根据上节中的簇中心寻找最近簇：

$$k^\star(z) = \arg\min_k \|z - \mu_{c,k}\|_2.$$

步骤 2：类内方差估计。 利用批内样本与簇中心的欧氏距离构造方差期望：

$$\hat{v}_c = \sum_{k=1}^{K} \pi_{c,k} \left(\frac{1}{|S_{c,k}^{(B)}|} \sum_{z \in S_{c,k}^{(B)}} \|z - \mu_{c,k}\|_2^2 \right), \tag{3}$$

其中 $S_{c,k}^{(B)} = \{z \in Z_B^{(c)} \mid k^\star(z) = k\}$，若为空集则忽略该簇。该量可被视为真实类条件分布 $p(z|y=c)$ 的经验二阶矩近似。

步骤 3：门控（Log-entropy）变换。 设定阈值 τ 与平滑项 ε，定义

$$\mathcal{R}_{\log}(c) = \max\left(0, \log(\hat{v}_c + \varepsilon) - \log(\tau + \varepsilon)\right), \tag{4}$$

或线性形式 $\mathcal{R}_{\text{lin}}(c) = \hat{v}_c$。最终 CEL 为

$$\mathcal{L}_{\text{CEL}} = \sum_{c=1}^{C} \beta_c \, \mathcal{R}(c), \tag{5}$$

其中 $\beta_c = \frac{|Z_B^{(c)}|}{|B|}$ 是当前 batch 的类别占比。

3.4 梯度与优化

总损失 $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{CEL}}$。为强调 CEL 仅用于约束客户端编码器，实践中采取“先对 $\mathcal{L}_{\text{CEL}}$ 反向传播捕获梯度，再清零梯度、对 $\mathcal{L}_{\text{CE}}$ 反传并与前者合并”的策略，使得 CEL 的梯度只作用于 $\boldsymbol{\theta}$。

3.5 特性小结

- 优点：统计更精细，能够捕捉多峰分布；对噪声注入、防御策略兼容性好。
- 缺点：需缓存全量 smashed data；聚类耗时且易受离群点影响；簇数 K 需调参。

4 gated-att 中的 CEL 计算流程

4.1 总体思路

gated-att 直接把“**可微注意力权重**”作为条件熵 surrogate，不再需要离线聚类。每个 batch 内，使用门控注意力对同类 smashed data 做加权聚合，计算加权方差，借此定义 CEL。注意力参数随训练迭代一同更新，实现端到端学习。

4.2 原始数据与批内处理

对于当前 batch B，考虑类别 c 的样本集合 $Z_B^{(c)}$，其大小记为 $m_c = |Z_B^{(c)}|$。引入门控注意力模块

$$\mathcal{A}_{\boldsymbol{\phi}} : \mathbb{R}^d \to (0,1), \quad \boldsymbol{\phi} = \{W_V, W_U, \boldsymbol{w}\},$$

其中

$$\boldsymbol{v}_i = \tanh(W_V z_i), \tag{6}$$

$$\boldsymbol{u}_i = \sigma(W_U z_i), \tag{7}$$

$$s_i = \boldsymbol{w}^\top (\boldsymbol{v}_i \odot \boldsymbol{u}_i), \tag{8}$$

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{m_c} \exp(s_j)}. \tag{9}$$

$\sigma(\cdot)$ 为 Sigmoid，$\odot$ 为逐元素乘法。

4.3 CEL 计算公式

加权统计量。

$$\bar{z}_c = \sum_{i=1}^{m_c} \alpha_i z_i, \tag{10}$$

$$v_c = \sum_{i=1}^{m_c} \alpha_i \|z_i - \bar{z}_c\|_2^2, \tag{11}$$

其中 α_i 满足 $\sum_i \alpha_i = 1$。

门控阈值。 同样提供线性或 log-entropy 两种形式：

$$\mathcal{R}_{\text{lin}}(c) = v_c, \tag{12}$$

$$\mathcal{R}_{\log}(c) = \max\left(0, \log(v_c + \varepsilon) - \log(\tau + \varepsilon)\right), \tag{13}$$

最终 CEL 为

$$\mathcal{L}_{\mathrm{CEL}} = \sum_{c=1}^{C} \beta_c \mathcal{R}(c), \qquad \beta_c = \frac{m_c}{|B|}. \tag{14}$$

Warm-up 策略。 为了避免初期注意力未收敛导致训练震荡，引入 warm-up 轮数 T_{warm}：在 $t \le T_{\mathrm{warm}}$ 的 epoch 内 $\mathcal{L}_{\mathrm{CEL}}$ 被禁用，仅训练分类损失；在 $t > T_{\mathrm{warm}}$ 时才逐渐打开 CEL，并按缩放系数 γ 调节其贡献：

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda\gamma\mathcal{L}_{\mathrm{CEL}}, \quad 0 < \gamma \le 1. \tag{15}$$

4.4 梯度与优化

由于 $\mathcal{L}_{\mathrm{CEL}}$ 完全可微，其梯度同时作用于编码器参数 $\boldsymbol{\theta}$ 与注意力参数 $\boldsymbol{\phi}$。当首次启用 CEL 时，将 $\boldsymbol{\phi}$ 加入优化器的参数组，其学习率、动量等超参与主干一致或单独设定。

4.5 特性小结

- 优点：端到端可微，无需存储全量 smashed data；对动态分布变化响应快。
- 缺点：注意力参数需额外调整；对 batch 内样本量较少的类别方差估计可能有偏，需要通过 ε 或正则项稳定。

5 步骤对比与时间/空间开销分析

表 1: 两种方案的关键差异一览

方面	`CEM-main`（聚类法）	`gated-att`（注意力法）
原始数据	全量 smashed data	当前 batch smashed data
统计方式	K-means / GMM 离线估计中心、方差	门控注意力实时计算加权均值与方差
是否缓存	需缓存所有 smashed data	无需缓存
额外参数	无，纯统计	注意力参数 $(W_V, W_U, \boldsymbol{w})$
CEL 形式	$\sum_c \pi_{c,k}\|z-\mu_{c,k}\|^2$ 的 log/线性门控	加权方差 v_c 的 log/线性门控
训练阶段	每轮结束后聚类，下一轮使用	每个 batch 实时计算
梯度流向	仅回传到编码器 $\boldsymbol{\theta}$	同时更新编码器 $\boldsymbol{\theta}$ 和注意力 $\boldsymbol{\phi}$
超参敏感性	对簇数 K、初始化敏感	对 warm-up、注意力维度敏感
算法复杂度	聚类 $O(NKd)$，内存 $O(Nd)$	每 batch $O(\|B\|d)$，内存 $O(\|B\|d)$

6 手工示例：三类二维特征的完整推导

为了更直观展示两种 CEL 的计算过程，下面构造一个小规模示例。设编码器输出为二维向量，存在三个类别，每类 2 条 smashed data：

$$
\begin{aligned}
Z^{(1)} &= \{(0.0, 0.0), (0.2, 0.1)\},\\
Z^{(2)} &= \{(1.0, 1.0), (1.2, 1.1)\},\\
Z^{(3)} &= \{(0.9, -0.9), (1.1, -1.0)\}.
\end{aligned}
$$

考虑 batch 恰好包含全部 6 条样本，label 均匀。取阈值 $\tau = 0.05$，平滑 $\varepsilon = 10^{-6}$，聚类簇数 $K = 1$。

6.1 CEM-main 计算步骤

离线聚类。 由于每类仅两点，且 $K=1$，聚类结果就是各自的样本均值：

$$\begin{aligned}\mu_{1,1} &= (0.1, 0.05),\\ \mu_{2,1} &= (1.1, 1.05),\\ \mu_{3,1} &= (1.0, -0.95).\end{aligned}$$

簇权重 $\pi_{c,1}=1$。方差估计：

$$\begin{aligned}v_{1,1} &= \frac{1}{2}\big(\|(0,0)-\mu_{1,1}\|_2^2 + \|(0.2,0.1)-\mu_{1,1}\|_2^2\big)\\ &= \frac{1}{2}\left(0.125^2 + 0.125^2\right) \approx 0.03125,\\ v_{2,1} &\approx 0.0100,\\ v_{3,1} &\approx 0.0100.\end{aligned}$$

批内方差。 由于 batch 与离线集合一致，再次计算也得到相同的 $\hat{v}_c = v_{c,1}$。

CEL 计算。 采用 log-entropy：

$$\begin{aligned}\mathcal{R}_{\log}(1) &= \max\big(0, \log(0.03125+\varepsilon) - \log(0.05+\varepsilon)\big) = 0,\\ \mathcal{R}_{\log}(2) &= 0,\\ \mathcal{R}_{\log}(3) &= 0.\end{aligned}$$

若使用线性形式，则

$$\mathcal{L}_{\mathrm{CEL}}^{\mathrm{lin}} = \sum_{c=1}^{3} \frac{2}{6} v_{c,1} \approx 0.0171.$$

梯度方向。 对类别 1 的第一个样本 $z_1=(0,0)$，梯度为

$$\frac{\partial \mathcal{L}_{\mathrm{CEL}}^{\mathrm{lin}}}{\partial z_1} = \frac{2}{6}\cdot 2(z_1-\mu_{1,1}) = \frac{2}{3}(-0.1, -0.05).$$

说明编码器参数将被调整，使类别 1 的样本更接近其中心。

6.2 gated-att 计算步骤

注意力权重（示例取等权）。 假设经过若干轮训练，注意力网络已学习到对称权重，即 $\alpha_i = \frac{1}{2}$。则

$$\bar{z}_c = \frac{1}{2} z_1^{(c)} + \frac{1}{2} z_2^{(c)} = \mu_{c,1}.$$

加权方差。

$$v_1 = \frac{1}{2}\|(0,0) - \mu_{1,1}\|_2^2 + \frac{1}{2}\|(0.2,0.1) - \mu_{1,1}\|_2^2 \approx 0.03125,$$
$$v_2 \approx 0.0100, \quad v_3 \approx 0.0100.$$

CEL。 与聚类法一致，线性形式 $\mathcal{L}_{\text{CEL}}^{\text{lin}} \approx 0.0171$。

注意力梯度。 若实际训练中 α_i 不是 $\frac{1}{2}$，CEL 的梯度将推动注意力网络向“抑制远离中心的样本”方向更新。例如若类别 1 中注意力权重 $\alpha_1 = 0.7, \alpha_2 = 0.3$，则

$$\bar{z}_1 = 0.7\, z_1 + 0.3\, z_2, \quad v_1 = 0.7\|z_1 - \bar{z}_1\|_2^2 + 0.3\|z_2 - \bar{z}_1\|_2^2.$$

若 z_1 离中心更远，梯度会增大注意力网络对 z_2 的权重，以减小 v_1。

6.3 示例结论

在上述简单示例中，两种方法给出的线性 CEL 数值相同。但在更复杂数据上，`CEM-main` 依赖历史聚类，可能捕捉到 batch 之外的全局结构；`gated-att` 则利用注意力实时对 batch 内样本加权。前者对批间一致性敏感，后者对注意力网络的训练稳定性敏感。

7 实践建议与调参要点

7.1 `CEM-main`

- **簇数** K：可按类别 smashed data 的多峰程度设定。经验上 CIFAR-10 使用 $K = 5$–10。
- **聚类频率**：可每轮聚类一次，也可每若干轮聚类以降低开销。
- **阈值** τ：与噪声注入强度 σ^2 协同调节，过小会导致梯度爆炸。

7.2 `gated-att`

- **Warm-up 轮数**：建议至少 5–10 轮，确保分类性能先收敛。
- **注意力隐层维度**：默认取 $d/4$ 或 128，过大易过拟合，过小表达不足。
- **缩放系数** γ：可从 0.1 起逐渐增大，观察对准确率与重建质量的影响。

8 结论

本文以数学视角详细梳理了 `CEM-main` 与 `gated-att` 在条件熵正则上的实现差异。二者在目标上一致，均致力于压缩 smashed data 的类内可分散度；但在具体实现中，一个依赖离线聚类、另一个借助门控注意力。根据实际场景的计算资源、数据规模、对实时性的要求，可灵活选择或进一步结合两者的优点（例如利用注意力初始化聚类中心，或在注意力损失中引入历史统计的先验项）。

参考阅读

- Xia et al., "Theoretical Insights in Model Inversion Robustness and Conditional Entropy Maximization for Collaborative Inference Systems," CVPR 2025.
- Ilse et al., "Attention-based Deep Multiple Instance Learning," ICML 2018.
- 常见模型反演攻击综述与开源实现（`GMI`, `BiDO` 等）。