Slot + Gated Cross 条件熵代理框架解析报告

自动生成

November 6, 2025

Abstract

本文面向项目根目录中的`SlotCrossAttentionCEM` 模块，系统梳理“slot + gated cross” 条件熵（Conditional Entropy Minimization, CEM）代理的设计思想、数学建模与训练流程。报告首先回顾CEM 在隐私防护/表示学习中的目标，再详细拆解Slot Attention、门控交叉注意力、混合槽统计与门控策略等关键组件，并解释这些设计如何替代原有的高斯混合模型（GMM）近似器来稳定地估计条件熵损失。最后我们分析该模块在训练循环中的反向传播路径、超参数选择与调试建议。

Contents

1 背景：条件熵最小化（CEM）

1.1 目标函数

在会员推理防御与判别式表示学习任务中，我们希望编码器产出的特征Z 尽可能与类别标签Y 对应且类内紧凑。条件熵

$$H(Z \mid Y) = \sum_{c} p(c)\, H(Z \mid Y = c) \tag{1}$$

刻画了在给定类别情况下特征分布的不确定性。若我们能最小化$H(Z \mid Y)$，就能抑制类内方差、削弱隐私攻击对样本细节的重构能力。

在实践中不可直接获得$p(z \mid y)$，常见近似是假设类条件分布近似高斯，此时

$$H(Z \mid Y = c) \approx \frac{1}{2} \log \det\bigl(2\pi e\, \Sigma_c\bigr), \tag{2}$$

其中Σ_c 为类别c 的协方差矩阵。传统方法通常使用KMeans/GMM 来拟合类别子簇，再估计方差或熵。然而在高维空间、批大小有限或者分布多模态时，这类方法容易数值不稳定或对初始化敏感。

1.2 设计动机

“slot + gated cross” 框架的设计出发点在于：

- 用Slot Attention 自适应地捕获类内多个潜在子模式（代替GMM 的硬聚类中心）。
- 通过Flamingo 风格的门控交叉注意力，让每个样本在共享槽（slot）上下文中提取增强特征，获得更稳定的局部统计量。
- 在方差估计上引入多级门控（维度门、SNR 门、槽质量门、类级门）与软阈处理，以抑制噪声方差并控制梯度爆炸。

2 框架总览

整体流程按类别拆分为三个阶段：

1. **Slot Attention 汇聚**：将同类样本嵌入$\{x_m\}_{m=1}^{M}$ 看成序列，经过多次竞争-更新获得一组共享槽$\{s_k\}_{k=1}^{S}$。
2. **门控交叉注意力增强**：以样本特征为查询、槽为键值，通过多头交叉注意力并配合门控残差，产生增强后的特征$\{\tilde{x}_m\}$。
3. **混合槽统计& 条件熵surrogate**：基于归一化相似度推导软责任r_{mk}，进而得到槽均值、槽方差，并在多级门控与阈值平滑下组合成类级的对数方差指标。

3 组件详解

3.1 Slot Attention 机制

Slot Attention 模块位于`model_training_paral_pruning.py` 第35–87 行。输入为形状$[B, N, D]$ 的序列，其中当前用法是$B = 1$、$N = M$（类内样本数）。每次迭代步骤如下：

$$k_n = W_k \,\mathrm{LN}(x_n), \quad v_n = W_v \,\mathrm{LN}(x_n), \tag{3}$$

$$q_k = W_q \,\mathrm{LN}(s_k^{(t-1)})/\sqrt{d}, \tag{4}$$

$$a_{nk} = \mathrm{softmax}_k\left(\frac{k_n^\top q_k}{\tau}\right), \qquad \tau = (d)^{1/4}, \tag{5}$$

$$u_k = \sum_n a_{nk} v_n, \tag{6}$$

$$s_k^{(t)} = \mathrm{GRU}\big(u_k, s_k^{(t-1)}\big) + \mathrm{MLP}\big(\mathrm{LN}(\mathrm{GRU}(\cdot))\big). \tag{7}$$

相较原始论文，这里额外引入$\tau = d^{1/4}$ 的温度来缓和对数似然值的尺度，避免梯度过大；同时层归一化保证不同批次的数值稳定。

3.2 门控交叉注意力

交叉注意力模块定义在第89–131 行。给定单个样本的特征x_m（视为$[1, D]$），以及共享槽$\{s_k\}$，流程为：

$$q = \mathrm{LN}_q(x_m), \quad K = \mathrm{LN}_{kv}(S), \quad V = \mathrm{LN}_{kv}(S), \tag{8}$$

$$\alpha = \mathrm{softmax}\left(\frac{qK^\top}{\sqrt{d}}\right), \tag{9}$$

$$h = \alpha V, \qquad o = W_o h, \tag{10}$$

$$y = x_m + \tanh(\alpha_{\mathrm{xattn}}) \cdot o, \tag{11}$$

$$y = y + \tanh(\alpha_{\mathrm{ffn}}) \cdot \mathrm{FFN}(\mathrm{LN}_{ff}(y)). \tag{12}$$

其中α_{xattn}、α_{ffn} 是可学习的门控系数，初始值较小使得网络可以平滑地从恒等映射过渡到使用注意力。该结构借鉴了Flamingo 模型的*Gated Cross-Attn Dense* 设计，使得注意力和前馈子层的贡献可被动态调节。

3.3 混合槽统计与条件熵surrogate

位于第179–275 行的是核心的CEM surrogate 计算。对每个类别c:

1. **样本归一化**：对特征做LayerNorm，缓解尺度差异导致的偏置。
2. **槽责任分配**：采用余弦相似度

$$s_{mk} = \frac{\langle \hat{x}_m, \hat{s}_k \rangle}{\|\hat{x}_m\| \cdot \|\hat{s}_k\|}, \qquad r_{mk} = \frac{\exp(\beta s_{mk})}{\sum_{k'} \exp(\beta s_{mk'})}, \tag{13}$$

   其中β（代码中的`assign_temp`）是可学习的温度，控制责任分布的尖锐程度。槽质量$w_k = \sum_m r_{mk}$ 用来度量该槽代表的样本量。

3. **槽均值与槽方差**:

$$\mu_k = \frac{1}{w_k} \sum_m r_{mk}\, \tilde{x}_m, \tag{14}$$

$$\sigma^2_{k,d} = \frac{1}{w_k} \sum_m r_{mk} \big(\tilde{x}_{m,d} - \mu_{k,d}\big)^2, \tag{15}$$

   并通过$\log \sigma^2_{k,d}$ 来后续构建熵surrogate。

4. **维度门控**：对每个槽的$\log \sigma^2$ 先经LayerNorm，再经过带Sigmoid 输出的MLP（第144–152 行）得到软门$g^{(\text{soft})}_{k,d}$，用于抑制噪声方差；同时计算信噪比

$$\text{SNR}_{k,d} = \frac{\sigma^2_{k,d}}{\mu^2_{k,d} + \varepsilon}, \tag{16}$$

   并通过$\text{sigmoid}(\alpha_{\text{snr}}(\text{SNR} - \tau_{\text{snr}}))$ 形成近似硬门$g^{(\text{hard})}_{k,d}$。

5. **软阈平滑**：相较于直接使用$\text{ReLU}(\log \sigma^2 - \log \sigma^2_{\text{thr}})$，这里改为

$$\phi_{k,d} = \frac{1}{\beta_{\text{sp}}} \log\left(1 + \exp\big(\beta_{\text{sp}}(\log \sigma^2_{k,d} - \log \sigma^2_{\text{thr}} - m)\big)\right), \tag{17}$$

   对应代码中的`softplus_beta` 与`margin_m`。该设计可以在阈值附近提供平滑梯度。

6. **槽权重门控**：槽质量经幂次γ（`slot_power`）强化后归一化:

$$\tilde{w}_k = \frac{(w_k/M)^\gamma}{\sum_{k'} (w_{k'}/M)^\gamma}, \tag{18}$$

   使得代表性较强的槽贡献更大。

7. **类级门控与聚合**：将维度门、SNR 门、阈平滑项相乘

$$\psi_{k,d} = g^{(\text{soft})}_{k,d} \cdot g^{(\text{hard})}_{k,d} \cdot \phi_{k,d}, \tag{19}$$

   然后先按槽权重求和，再对维度求平均得到类级surrogate $\hat{L}_c$:

$$\hat{L}_c = \frac{1}{D} \sum_d \sum_k \tilde{w}_k\, \psi_{k,d}. \tag{20}$$

为了避免小批次导致估计偏差，再用类别样本占比$p_c = M/B$ 进入Sigmoid 门

$$g_c = \sigma\big(a(p_c - b)\big), \tag{21}$$

（`class_gate_a`, `class_gate_b`），最终类级贡献为$g_c \cdot \hat{L}_c$。

对所有类别求加权平均（权重即p_c）即可得到

$$\text{rob\_loss} = \sum_c p_c \, g_c \, \hat{L}_c, \qquad \text{intra\_mse} = \sum_c p_c \, \text{MSE}_c, \tag{22}$$

其中MSE_c 仅用于日志监控。

4 门控策略与稳定性设计

4.1 早期关断机制

在训练初期模型尚未收敛时，方差估计极易震荡。代码通过维护滑动平均的门值，并在以下任一条件成立时将rob_loss 强制置零：

- 模块调用次数$\leq$ `early_shut_steps`。
- SNR 硬门平均值大于`early_hard_thresh`。
- 软门平均值大于`early_gate_thresh`。

这确保早期梯度不会把编码器推向坏的局部最优，同时仍保留计算图以便后续渐进启用。

4.2 可学习温度与阈值

多处门控参数都是可学习的标量，允许网络在训练过程中自动调节：

- β：责任分布温度，控制槽的“硬”程度。
- γ：槽质量幂，决定是否强调主槽。
- SNR 阈值与斜率$\tau_{\text{snr}}, \alpha_{\text{snr}}$：自动判别低信噪比维度。
- Softplus 斜率与边际(β_{sp}, m)：稳定阈值附近的梯度。
- 类级门(a, b)：基于批量样本比例调节贡献，避免样本数极小的类别主导损失。

5 与CEM 训练流程的结合

5.1 前向阶段

主训练循环在`train_target_step` 中实现（第1240–1501 行）。当满足

$$\texttt{use\_attention\_cem} \wedge \neg \texttt{random\_ini\_centers} \wedge \lambda > 0 \wedge \texttt{epoch} > \texttt{warmup}$$

时，会将特征展平并送入注意力版CEM，得到rob_loss 与intra_class_mse。随后将rob_loss 乘以缩放系数（默认0.25）以控制梯度量级。

5.2 反向与梯度合成

训练器先对rob_loss 做一次反向传播，仅保留编码器和注意力模块的梯度快照，然后清零优化器梯度；接着对分类损失L_{CE} 反向，最终以

$$\nabla_{\theta_f} = \nabla_{\theta_f} L_{\mathrm{CE}} + \lambda \cdot w_{\mathrm{lr}} \nabla_{\theta_f} \mathrm{rob\_loss} \tag{23}$$

的形式合并梯度，其中w_{lr} 是基于学习率调度器的缩放项。注意力模块参数则直接累加两次反向的梯度。该做法相当于执行一次“显式梯度加权”，避免在总损失中直接拼接两个项导致训练初期不稳定。

6 实现与调优建议

6.1 超参数与默认值

- 槽数量$S = 8$（可根据类内模式复杂度调节）。
- Slot Attention 迭代次数$T = 3$，折中考虑效率与表示力。
- 责任温度、槽幂、SNR 阈值、Softplus 斜率等均为可学习参数，训练中会自动调整。
- 训练初期使用`attention_warmup_epochs`=3 暂停CEM 梯度，待编码器初步收敛后再启用。
- rob_loss 的缩放`attention_loss_scale`=0.25，可以视情况放大以增强防护，也可减小以保证主任务精度。

6.2 数值稳定性心得

- 在高维特征上，LayerNorm 与温度缩放是必要的，否则余弦相似度与对数方差容易溢出。
- Softplus 边际m 能防止σ^2 略低于阈值时出现梯度突变，建议与`var_threshold` 联动调节。
- 若批次类别数量过少，可适当放宽`early_shut` 条件或降低类级门a，否则CEM 长时间处于关闭状态。

7 结论

Slot + Gated Cross 框架以模块化方式重写了CEM 近似器：Slot Attention 捕获多模态子簇，门控交叉注意力在共享上下文中重投影样本，随后利用多重门控的混合槽统计稳定地估计类内对数方差，从而提供光滑、可控的条件熵surrogate。通过分阶段反向与梯度合并，该模块实现了与主任务的解耦训练，同时保留对编码器的正则化力度。实务中建议结合warmup、门控参数监控与梯度缩放来调节防护-精度平衡。