Slot + Gated Cross CEM 易懂手册（v6）

自动生成

November 10, 2025

Abstract

这份v6 版手册对应根目录的`SlotCrossAttentionCEM`，配套流程图`slot_gated_cem_flowchart_orig`
我们保持原有流程（Slot Attention → Gated Cross Attention → 混合槽统计），
但把解释方式改成“第一手就能看懂”的结构，方便直接拿去做PPT 或讲稿。

Contents

1 一句话背景

- **目标**：让同一类别的特征在中间层更紧凑，降低条件熵，减小被攻击者还原的风险。
- **方法**：用“队长+ 向队长学习”来概括类内多样性，再用多重门控判定哪些方差值得惩罚。
- **输出**：一个正则损失`rob_loss`（用于反向）和一个监控指标`intra_mse`。

2 总览：四个阶段，像排班表

1. **阶段1：Slot Attention 挑队长**
   同类样本站成一排，多个slots（队长）和大家多轮互动后，成为几个子簇代表。
2. **阶段2：Gated Cross Attention 向队长取经**
   每个样本根据注意力权重向队长借经验，门控残差确保逐步引入新信息。
3. **阶段3：混合槽统计判断“是否稳”**
   用余弦相似度分配责任，计算加权均值/方差，再通过维度门、SNR 门、Softplus 阈值和槽权重过滤噪声。
4. **阶段4：类级聚合并输出损失**
   对每类得到的结果乘以样本占比，合成为`rob_loss`；同时记录`MSE` 曲线。

3 阶段细节

3.1 1. Slot Attention（队长竞选）

- **输入形状**：$[B = 1, N = M, D]$，一次只处理某个类别的样本。
- **流程**：
  1. LayerNorm 后映射成Key/Value。
  2. slots 从可学习的高斯分布采样，避免完全一样。
  3. 多轮循环：
     - 用slots 发出Query，和样本Key 点积→ softmax 责任。
     - 按责任加权Value，得到更新量。
     - 用GRU 和MLP 更新slots（保留记忆+ 引入新信息）。
- **额外细节**：点积除以$(d)^{1/4}$ 防止梯度爆炸，LayerNorm 让不同batch 可比。

3.2 2. Gated Cross Attention（向队长取经）

- **输入**：原样本特征x_m 和上一步的slots。
- **流程**：
  1. 样本做Query，slots 做Key/Value。
  2. 计算注意力后得到增强特征。
  3. 通过$\tanh(\alpha_{\text{xattn}})$ 门控残差加入原特征。
  4. 再通过Pre-LN + FFN + $\tanh(\alpha_{\text{ffn}})$ 的门控残差。
- **意义**：门控让模型一开始保持原状，训练稳定后再逐步依赖注意力。

3.3 3. 混合槽统计（判断是否松散）

1. **责任分配**：增强后的样本与slots 做余弦相似度，softmax 得到r_{mk}。
2. **加权统计**：用r_{mk} 计算每个槽的加权均值μ_s、方差σ_s^2。
3. **三种门控**：
   - **维度软门**：LayerNorm + MLP + Sigmoid，抑制噪声维度。
   - **SNR 硬门**：$\sigma^2/(\mu^2 + \varepsilon)$，信噪比低的压低。
   - **Softplus 阈值**：对$\log \sigma^2$ 做平滑阈值，避免ReLU 的断点。
4. **槽权重**：根据slot mass 调整权重，代表性强的槽更重要。
5. **类级聚合**：门控项相乘→ 按槽权重求和→ 对维度取平均，得到L_c；再乘以类级Sigmoid 门（batch 样本少则权重低）。

4 训练中的使用姿势

- **触发条件**：$\lambda > 0$、非随机中心、epoch 超过`attention_warmup_epochs`。
- **梯度流程**：
  1. rob_loss 先反向，保存编码器/注意力梯度。
  2. 清梯度，再对交叉熵反向。
  3. 按学习率缩放& λ 加回rob_loss 的梯度。
- **保护机制**：早期关断、NaN/Inf 自动置零，确保不会把主任务拉崩。

5 可调旋钮与排障

6.1 常用参数

- `num_slots`：默认8，可按类内复杂度调。
- `num_iterations`：默认3，越大越精细但越慢。
- `attention_loss_scale`：默认0.25，控制正则强度。
- `var_threshold`：越小越严格，越大越宽松。

6.2 常见问题

- **rob_loss 一直是0?**
  可能还在warmup，或门值触发了早期关断；看日志里的`[CEM-GATE]` 提示。
- **rob_loss 过大导致精度掉**?
  先减小`attention_loss_scale`；也可以调大`var_threshold`。
- **NaN/Inf**?
  检查输入特征是否已有NaN；或调大Softplus 边际、eps。

6 拿着这份文档做什么?

- 将本手册和`slot_gated_cem_flowchart_original.v6.pdf` 一起做成PPT。
- 在组会上按"阶段概述→ 细节→ 调参"三段讲法分享。
- 如需英文版或更技术化版本，可在此基础上补充数学公式。