Gated Attention 条件熵代理框架解析报告 (Original)

自动生成

November 11, 2025

Abstract

本文面向`gated-att/` 项目中的`GatedAttentionCEM` 模块，系统梳理门控注意力替代GMM/KMeans 的条件熵代理。内容包括目标动机、GatedAttentionPooling 的细节、加权统计与阈值化流程，以及在主训练循环中的集成方式。该版本保持原报告的数学表达，便于与流程图联动展示。

Contents

1 背景与动机

1.1 条件熵最小化

类似Slot+Cross 框架，这里希望压缩类内特征分布，使条件熵$H(Z \mid Y)$ 降低，以削弱攻击者在割点上的重建能力。传统GMM/KMeans 代理在高维/小批量情况下不稳定，因此采用门控注意力直接学习类内加权统计。

1.2 为什么用Gated Attention

- 结构更轻：单层gated attention 即可得到权重，不需要slots 或多轮竞争，适合设备受限或想快速验证思路的场景。
- 仍可学到“关键样本”：注意力权重会偏向代表性强的样本，用加权方差来衡量类内散度。
- 直接替换GMM：把attention 权重理解为“软聚簇概率”，允许端到端训练。

2 组件详解

2.1 GatedAttentionPooling（§17–44 行）

1. 输入为同类样本矩阵$X_c \in \mathbb{R}^{M \times D}$。

2. 通过两条支路：

$$V = \tanh(W_V X_c), \qquad U = \sigma(W_U X_c),$$

其中tanh 捕捉方向，σ 充当门控。

3. 两条支路逐元素相乘后映射到标量logits：

$$\ell = W_w(V \odot U) \in \mathbb{R}^{M\times 1}.$$

4. 沿样本维度做softmax 得到注意力权重$a = \mathrm{softmax}(\ell)$，满足$\sum_m a_m = 1$。

权重a 中较大的样本被认为是类内关键信息，后续加权均值/方差都以此为系数。

2.2 GatedAttentionCEM（§46–118 行）

对每个类c：

1. LayerNorm：$\hat{X}_c = \mathrm{LayerNorm}(X_c)$，对齐尺度。
2. 通过pooling 得到注意力a。
3. 加权均值：

$$\mu_c = \sum_m a_m \hat{x}_m.$$

4. 加权方差：

$$\sigma_c^2 = \sum_m a_m (\hat{x}_m - \mu_c)^2,$$

并以eps 下界避免$\log 0$。

5. 阈值化log-variance：

$$\log \sigma_c^2 = \log(\sigma_c^2 + \gamma), \qquad \mathrm{ce\_sur} = \max\big(0,\ \log \sigma_c^2 - \log \sigma_{\mathrm{thr}}^2\big),$$

其中$\sigma_{\mathrm{thr}}^2 = \max(\texttt{var_thr} \cdot \texttt{reg_strength}^2, 10^{-8}) + \gamma$。

6. 类级输出$L_c = \mathrm{mean}(\mathrm{ce_sur})$，并记录MSE 作为监控指标。
7. 按样本占比$w_c = M/B$ 加权，加到总的rob_loss 与$intra_mse$ 中。

3 训练循环中的集成

- 在‘train$_t arget_s tep$‘−fi $> 0\Delta epoch > warmup\Delta\hat{}$:−fi1æΛ(‘$GatedAttentionCEM$‘‘$rob_l oss$‘Φ‘$model_l$ $742 - 902$‘ΨΘ
- 输出的‘rob$_l oss$‘X‘$attention_l oss_s cale$‘$vH_\mathrm{J} X$fiffi;ΦIΨΦ$_1$Θ
- NaN/Inf 的结果会清零，保证主干训练不被破坏。

4 数值稳定性与调参

- `hidden_dim`：注意力MLP 的隐藏维度，自动取$[64, 512]$。
- `var_threshold`, `reg_strength`：决定log-variance 阈值；阈值越大越宽松。
- `attention_loss_scale`：控制正则梯度大小；训练稳定后可调大。
- LayerNorm + eps 防止数值问题，类内样本不足时则跳过该类。

5 结论

Gated Attention CEM 用单层gated pooling + 阈值化log-variance 的方式，替代GMM/KMeans 作为类内条件熵surrogate。结构轻量、易部署，适合在算力有限或类内单模态场景下优先使用。该报告与流程图（‘gated$_a ttention_c em_f lowchart_o riginal.v6.pdf$‘$\Psi MW$fi($U$:Θ