Attention-CEM 防御框架总览

Slot+Gated Cross Attention vs. Gated Attention Pooling

项目阶段性汇报

Attention Privacy Project

组会交流

演示地图

为什么

- 协同推理的隐私风险
- CEM 框架基本思路

Gated Attention Pooling

- 方案动机与流程
- 条件熵代理公式

Slot+Gated Cross Attn

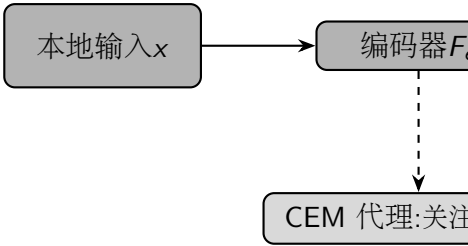
- 模块结构图
- 数学细节与门控
- 训练集成

总结与计划

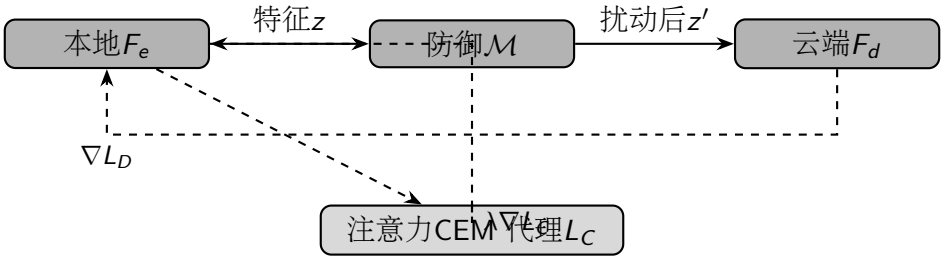
- 两者并列对比
- 实践建议与风险点
- 下一步重点

协同推理与CEM 目标

- **协同推理**:本地编码器F_e 产生中间特征z，云端解码器F_d 完成预测。
- **隐私问题**:中间特征易被模型反演攻击恢复输入。
- **CEM 思路**:最大化条件熵$H(x|z)$，提高攻击者的最优重建误差ξ。
- 原始代码使用**KMeans/GMM** 近似条件熵，下文两套注意力框架即替换该近似器。



训练循环中的注意力CEM



- : “”
- :

SlotAttention

1. LayerNorm K, V
2. μ, σ S slot
3. :
   - slot$\rightarrow$ Q
   - $r = \text{softmax}(KQ^\top)$ ϵ
   - GRU+MLP slot

- slot $S = 8$
- $T = 3$
- :slot extunderscore $\text{dim}^{-1/4}$

Gated Cross Attention

- :

$$y = q + \tanh(\alpha_{\text{attn}}) \cdot \mathrm{CrossAttn}(q, s), \quad y = y + \tanh(\alpha_{\text{ffn}}) \cdot \mathrm{FFN}(\mathrm{LN}(y)).$$

- :slots KV
- (0.1)/FFN
- LayerNorm

+ CEM

1. **Slot** :$r_{ms} = \mathrm{softmax}(\beta \cdot \mathrm{sim}(x_m, s_s))$ μ_s, σ_s^2
2. **Per-dim Gate**:LayerNorm$(\log \sigma_s^2) \rightarrow$MLP$\rightarrow$Sigmoid
3. **SNR Gate**:$g_{\mathrm{snr}} = \sigma(\kappa(\sigma^2/(\mu^2 + \epsilon) - \tau_{\mathrm{snr}}))$
4. **Softplus Margin**:$L_{\mathrm{base}} = \frac{1}{\beta'} \log(1 + e^{\beta'(\log \sigma^2 - \log \tau - m)})$
5. **Slot Mass Gate**:$(\mathrm{mass}/M)^{\gamma}$ slot
6. **Class Gate**:$g_{\mathrm{class}} = \sigma(a(M/B - b))$
7. **Early Shutoff**: 100 0

1. Warmup:‘self.attention_warmup_epochs $= 3$‘
2. ‘SlotCrossAttentionCEM‘
3. ‘rob_loss‘ / L_D
4. CEM λ ‘attention_loss_scale‘
5. DropoutARL

- Slot batch
- Gated Attention Pooling :

$$a_m = \frac{\exp\left(w^\top [\tanh(Vx_m) \odot \sigma(Ux_m)]\right)}{\sum_j \exp(\cdot)},$$

$$\mu = \sum_m a_m x_m, \qquad \sigma^2 = \sum_m a_m (x_m - \mu)^2,$$

$$L_C = \max\{0,\ \log(\sigma^2 + \gamma) - \log(\tau)\},$$

$$\tau = \texttt{var\_threshold} \cdot \texttt{reg\_strength}^2 + \gamma.$$

- + softmax
- LayerNorm
- slot

- Warmup 5 epoch softmax
-
- ‘attention_loss_scale‘ CEM 0.25
- Slot / Dropout / ARL
- batch

	Slot + Gated Cross Attn	**Gated Attention Pooling**
/	slot + cross-attn + early shutoff GRU/slot/ slot extunderscore power-class extunderscore gate	“” LayerNorm + softmax MLP batch var extunderscore threshold loss

- Warmup ‘current$_{e}poch$‘
- ‘rob$_{l}oss$‘$MSENaN/Inf$
- Slot early shutoff

- Slot : ‘slot$_{p}ower$“$class_{g}ate_{a}/b$“$attention_{l}oss_{s}cale$‘
- Gated : ‘reg$_{s}trength$“$var_{t}hreshold$‘$loss$

- CIFAR-10/100FaceScrubTinyImageNet
- GMM MIA MSE/SSIM
- :Gated pooling $\rightarrow$ Slot
- early shutoffclass gate -