$X = (x_t)_{t=1}^{T}$: RGB or flow features (with a pre-trained-model)

[attention-based action localization. problem, target is to predict the frame attention. λ.]

MAP: $\max_{\lambda \in [0,1]} \log P(\lambda \mid X, y)$ $P(\lambda \mid X, y)$ 给定 X, y unknown ~~probabili~~ probability distribution of λ:

maximum a posteriori.

~~frame level labels (ground truth of λ).~~

$$= \log \frac{P(X,y \mid \lambda) P(\lambda)}{P(X,y)}$$

$$\log P(\lambda \mid X, y) = \log P(X,y \mid \lambda) + \log P(\lambda) - \log P(X,y)$$

$$= \log \frac{P(X,y,\lambda)}{P(\lambda)} + \log P(\lambda) - \log P(X,y)$$

~~$= \log P(y, X, \lambda)$~~

$$= \log \frac{P(X,y,\lambda)}{P(X,\lambda)} \cdot \frac{P(X,\lambda)}{P(\lambda)} + \log P(\lambda) - \log P(X,y)$$

$$= \log P(\lambda \mid X, y) + \log P(X \mid \lambda) + \log P(\lambda) - \log P(X,y)$$

$$\propto \log P(y \mid X, \lambda) + \log P(X \mid \lambda)$$

prefers λ. ~~生成模~~ 判别模型 classification. 从λ 准确预测 帧的特征.

前景与背景 施加不同的关注

$P(X \mid \lambda)$ 用 generative model. 估计 使得模型精准重建特征

discriminative attention and the generative mo

/

Discriminative attention modeling.

前景 attention λ as weight;

前景 $x_{fg} = \frac{\sum_{t=1}^{T} \lambda_t x_t}{\sum_{t=1}^{T} \lambda_t}$.

背景 $x_{bg} = \frac{\sum_{t=1}^{T} (1-\lambda_t) x_t}{\sum_{t=1}^{T} (1-\lambda_t)}$

$L_d = L_{fg} + \alpha L_{bg}$.

$L_d = L_{fg} + \alpha L_{bg} = -\log P_\theta(y|x_{fg}) - \alpha \log P_\theta(o|x_{bg})$.

最小化 最大化

优化 λ.

Generative Attention modeling

对 用CVAE对 不同帧特征分布建模. 区分前景与背景

VAE VAE推导

$L = \log P(x)$

$= \sum_z q(z|x) \log P(x)$

$= \sum_z q(z|x) \log \left(\frac{P(z,x)}{P(z|x)}\right)$

$= \sum_z q(z|x) \log \left(\frac{P(z,x)}{q(z|x)} \frac{q(z|x)}{P(z|x)}\right)$

$= \sum_z q(z|x) \log \left(\frac{P(z,x)}{q(z|x)}\right) + \sum_z q(z|x) \log \left(\frac{q(z|x)}{P(z|x)}\right)$ 散度

$= L^V + D_{KL}(q(z|x) \| P(z|x))$

$L^V = \sum_z q(z|x) \log \frac{P(z,x)}{q(z|x)}$

$= \sum_z q(z|x) \log \frac{P(x|z)P(z)}{q(z|x)}$

$= \sum_z q(z|x) \log \frac{P(z)}{q(z|x)} + \sum_z q(z|x) \log P(x|z)$

$= -D_{KL}(q(z|x) \| P(z)) + E_{q(z|x)}[\log P(x|z)]$

散度

$P(X|\lambda) = \prod_{t=1}^{T} P(X_t|\lambda_t)$.

$P_\psi(X_t|\lambda_t) \equiv E_{P_\psi(z_t|\lambda_t)}[P_\psi(X_t|\lambda_t, z_t)]$

代入 L^V 得到

$L^V_{CVAE} = -E_{q_\phi(z_t|X_t,\lambda_t)} \log P_\psi(X_t|\lambda_t, z_t)$

$+ \beta \cdot KL(q_\phi(z_t|X_t,\lambda_t) \| P_\psi(z_t|\lambda_t))$

$\approx -\frac{1}{L}\sum_{l=1}^{L} (\log p_\psi(X_t|\lambda_t, z_t^{l}$ 中

$P_\psi(X_t|\lambda_t, z_t) = N(X_t|f_\psi(\lambda_t, z_t), \sigma^2 * I)$

$P_\psi(z_t|\lambda_t) = N(z_t|r\lambda_t \cdot 1, I)$

$q_\phi(z_t|X_t,\lambda_t) = N(z_t|\mu_\phi, \Sigma_\phi)$ ← outputs of encoder $f_\phi(X_t,\lambda_t)$

$L_{re} = -\sum_{t=1}^{T} \log\{E_{P_\psi(z_t|\lambda_t)}[P_\psi(X_t|\lambda_t, z_t)]\} \approx \sum_{t=1}^{T} \log\left(\frac{1}{L}\sum_{l=1}^{L} P_\psi(X_t|\lambda_t, z_t^{l})\right)$

$L=1$ $L_{re} = -\sum_{t=1}^{T} \log P_\psi(X_t|\lambda_t, z_t) \propto \sum_{t=1}^{T} \|X_t - f_\psi(\lambda_t, z_t)\|^2$.

用之前 Discriminative attention Modeling

优化的 X. 来更新 CVAE 的时候生成 $P(X|\lambda)$.

L_{CVAE} 学习参数 L_{re} 更新入.

$$\hat{\lambda}_t^{fg} = G(\sigma_s) * \frac{\exp w_c^T x_t}{\sum_{c=0}^{C} \exp w_c^T x_t} \cdot \quad \hat{\lambda}_t^{bg} = G(\sigma_s) * \frac{\sum_{c=1}^{C} \exp w_c^T x_t}{\sum_{c=0}^{C} \exp w_c^T x_t}$$

$$L_{guide} = \frac{1}{T} \sum_{t=1}^{T} |\lambda_t - \hat{\lambda}_t^{fg}| + |\lambda_t - \hat{\lambda}_t^{bg}|$$

1. update attention and classification modules. with loss

$$L = L_d + \gamma_1 L_{re} + \gamma_2 L_{guide}$$

where γ_1 γ_2 denote the hyper-parameters

2. update CVAE with loss L_{CVAE}

重建 representation.

$P(X|\lambda)$ 就是具体那帧是动作

↓

更新参数通过 $P(X|\lambda)$

自底向上方法输入数据产生每一帧注意力 并通过注意力加权.

的特征. 训练分类模型

轻易得知识别能力

CVAE 用以注意力为条件的来模拟框架特征分布