可验证边界重证明

原文定理

来源于SAFER：A Structure-free Approach for certified robust

符号：分类器f，输入样本X，X'是对抗候选样本来自于S_X，Z是对应的扰动样本来自于P_x，注意两者的差别，Y是标签集合，Π_X是样本X的扰动样本的分布。

原文内容：平滑分类器f^{RS}是鲁棒的当$y = f^{RS}(X')$对于任何$X' \in S_X$，其中y是真实标签，S_X是加入同义词扰动的样本集合，$f^{RS}(X) = \underset{c \in Y}{argmax} P_{Z \sim \Pi_X}(f(Z) = c)$，$g^{RS}(X,c) = P_{Z\sim \Pi_X}(f(Z) = c)$，注意$\Pi_X$和$S_X$的扰动方法是不同的，$S_X$是获取词向量空间余弦相似度大于0.8的同义词集合，Π_X的扰动方法是来自于扰动集合P_x，扰动集合是生成于同义词集合，每个单词x的扰动集合是通过路径$x_1, x_2, \dots, x_n$，相邻的两个单词是同义词，则x_1, x_n在同一个扰动集合中，扰动集合大小有阈值K=100，超过阈值的集合取前K个余弦相似度最相近的单词。鲁棒的一个充分条件是

$$\underset{X' \in S_X}{min}\, g^{RS}(X', y) \geq \underset{X' \in S_X}{max}\, g^{RS}(X', c), \forall c \neq y$$

两个值相当于对抗样本的y标签的下界和c标签的上界，因此，关键步骤为计算两个边界。

可验证上界/可验证下界定理：假设扰动集合P_x被构造为$|P_x| = |P_{x'}|$，对于每个一个单词x和它的同义词$x' \in S_x$。定义

$$q_x = \underset{x' \in S_x}{min} |P_x \cap P_{x'}| / |P_x|$$

q_x体现了两个不同扰动集合之间的重叠程度，1为完全重叠，0为不重叠。对于一个给定的样本$X = x_1, x_2, \dots, x_n$，我们按照q_x从小到大排序$q_{x_{i1}} \leq q_{x_{i2}} \leq \dots \leq q_{x_{iL}}$，则

$$\underset{X' \in S_X}{min}\, g^{RS}(X', y) \geq max(g^{RS}(X, c) - q_X, 0)$$

$$\underset{X' \in S_X}{max}\, g^{RS}(X', c) \leq min(g^{RS}(X, c) + q_X, 1)$$

其中，$q_X = 1 - \Pi_{j=1}^{R} q_{x_{ij}}$，扰动R个单词。

命题：对于任何样本X和它的标签y，我们定义$y_B = arg \underset{c \in Y, c \neq y}{max}\, g^{RS}(X, c)$，我们能够证明$f(X') = f(X) = y$对于任意的$X' \in S_X$当

$$\Delta_X = g^{RS}(X, y) - g^{RS}(X, y_B) - 2q_X > 0$$

因此验证模型预测是否一致只要验证Δ_X是否是正的即可，可以通过蒙特卡洛采样估计。

我的定理

宽松

鲁棒的一个更严格的充分条件是：

$$\underset{X' \in S_X}{min}\, g^{RS}(X', c_A) - g^{RS}(X', c_B) > 0$$

命题：$\Delta_X = g^{RS}(X, y) - g^{RS}(X, y_B) > 0$即可证明$f(X') = f(X) = y$。

证明：

改 \mathcal{H}

定义$h \in H_{[0,1]}$，为将X映射到[0,1]的函数，（这里用到的h相当于映射X到某类别c的概率函数），定义$\Pi_X[h] = E_{Z \sim \Pi_X}[h(Z)]$，（相当于h的随机平滑版本）

$$\underset{X' \in S_X}{min}\, g^{RS}(X', c_A) - g^{RS}(X', c_B)$$

$$= \underset{h_A, h_B \in H_{[0,1]}}{min}\ \underset{X' \in S_X}{min} \Pi_{X'}[h_A] - \Pi_{X'}[h_B], st. \Pi_X[h_A] = g^{RS}(x, c_A), \Pi_X[h_B] = g^{RS}(x, c_B), \Pi_{X'}[h_A] + \Pi_{X'}[h_B] \leq 1$$

$$= \underset{X' \in S_X}{min}\ \underset{h_A, h_B \in H_{[0,1]}}{min}\ \underset{\lambda_1, \lambda_2, \lambda_3 \in R}{max} \Pi_{X'}[h_A] - \Pi_{X'}[h_B] - \lambda_1 \Pi_X[h_A] + \lambda_1 p_A - \lambda_2 \Pi_X[h_B] + \lambda_2 p_B + \lambda_3 \Pi_{X'}[h_A] + \lambda_3 \Pi_{X'}[h_B] - \lambda_3$$

$$where, p_A = g^{RS}(X, c_A), p_B = g^{RS}(X, c_B)$$

$$\geq \underset{\lambda_1, \lambda_2, \lambda_3 \in R}{max}\ \underset{X' \in S_X}{min}\ \underset{h_A, h_B \in H_{[0,1]}}{min} \int_Z h_A(Z) d\Pi_{X'}(Z) - \int_Z h_B(Z) d\Pi_{X'}(Z) - \backslash$$

$$\lambda_1 \int_Z h_A(Z) d\Pi_X(Z) - \lambda_2 \int_Z h_B(Z) d\Pi_X(Z) + \lambda_3 \int_Z h_A(Z) d\Pi_{X'}(Z) + \lambda_3 \int_Z h_B(Z) d\Pi_{X'}(Z) + \lambda_1 p_A + \lambda_2 p_B - \lambda_3$$

$$= \underset{\lambda_1, \lambda_2, \lambda_3 \in R}{max}\ \underset{X' \in S_X}{min}\ \underset{h_A, h_B \in H_{[0,1]}}{min} \int_Z h_A(Z)(d\Pi_{X'}(Z) - \lambda_1 d\Pi_X(Z) + \lambda_3 d\Pi_{X'}(Z)) + \int_Z h_B(z)(-d\Pi_{X'}(Z) - \lambda_2 d\Pi_X(Z) + \lambda_3 d\Pi_{X'}(Z)) + \lambda_1 p_A + \lambda_2 p_B -$$

$$= \underset{\lambda_1, \lambda_2, \lambda_3 \in R}{max}\ \underset{X' \in S_X}{min} - \int_Z (\lambda_1 d\Pi_X(Z) - (\lambda_3 + 1) d\Pi_{X'}(Z))_+ - \int_Z (\lambda_2 d\Pi_X(Z) - (\lambda_3 - 1) d\Pi_{X'}(Z))_+ + \lambda_1 p_A + \lambda_2 p_B - \lambda_3$$

$$where, (x)_+ = max(x, 0)$$

abort

现在我们推导$\int_Z (\lambda_1 d\Pi_X(Z) - (\lambda_3 + 1) d\Pi_{X'}(Z))_+$和$\int_Z (\lambda_2 d\Pi_X(Z) - (\lambda_3 - 1) d\Pi_{X'}(Z))_+$的形式。

定义$n_x = |P_x|, n_{x'} = |p_{x'}|, n_{x,x'} = |P_x \cap P_{x'}|$

$$\therefore \int_Z (\lambda_1 d\Pi_X(Z) - (\lambda_3 + 1) d\Pi_{X'}(Z))_+$$

$= \sum_{X' \in P_X \cap P_{X'}} (\lambda_1 |P_X|^{-1} - (\lambda_3+1)|P_{X'}|^{-1})_+ + \lambda_1 \sum_{X' \in P_X - P_X} |P_X|^{-1}$

$= |P_X \cap P_{X'}|(\lambda_1 |P_X|^{-1} - (\lambda_3+1)|P_{X'}|^{-1})_+ + \lambda_1 |P_X - P_{X'}||P_X|^{-1}$

$\because |P_X \cap P_{X'}| = \Pi_{i=1}^{L} n_{x,x'}$

$|P_X| = \Pi_{i=1}^{L} n_x$

$|P_{X'}| = \Pi_{i=1}^{L} n_{x'_i}$

$|P_X - P_{X'}| = |P_X| - |P_{X'} \cap P_X| = \Pi_{i=1}^{L} n_{x_i} - \Pi_{i=1}^{L} n_{x_i,x'_i}$

$\therefore |P_X - P_{X'}||P_X|^{-1} = \frac{\Pi_{i=1}^{L} n_{x_i} - \Pi_{i=1}^{L} n_{x_i,x'_i}}{\Pi_{i=1}^{L} n_{x_i}} = 1 - \Pi_{i=1,x_i \neq x'_i}^{L} \frac{n_{x_i,x'_i}}{n_{x_i}}$

$\therefore (\lambda_1 |P_X|^{-1} - (\lambda_3+1)|P_{X'}|^{-1})_+$

$= (\lambda_1 \Pi_{i=1}^{L} n_{x_i}^{-1} - (\lambda_3+1)\Pi_{i=1}^{L} n_{x'_i}^{-1})_+$

$= \Pi_{i=1,x_i=x'_i}^{L} n_{x_i}^{-1} (\lambda_1 \Pi_{i=1,x_i \neq x'_i}^{L} n_{x_i}^{-1} - (\lambda_3+1)\Pi_{i=1,x_i \neq x'_i}^{L} n_{x'_i}^{-1})_+$

$\therefore |P_X \cap P_{X'}|(\lambda_1 |P_X|^{-1} - (\lambda_3+1)|P_{X'}|^{-1})_+$

$= \Pi_{i=1}^{L} n_{x_i,x'_i} (\lambda_1 |P_X|^{-1} - (\lambda_3+1)|P_{X'}|^{-1})_+$

$= \Pi_{i=1,x_i=x'_i}^{L} n_{x_i} \Pi_{i=1,x_i \neq x'_i}^{L} n_{x_i,x'_i} (\lambda_1 |P_X|^{-1} - (\lambda_3+1)|P_{X'}|^{-1})_+$

$= \Pi_{i=1,x_i \neq x'_i}^{L} n_{x_i,x'_i} (\lambda_1 \Pi_{i=1,x_i \neq x'_i}^{L} n_{x_i}^{-1} - (\lambda_3+1)\Pi_{i=1,x_i \neq x'_i}^{L} n_{x'_i}^{-1})_+$

$= \Pi_{i=1,x_i \neq x'_i}^{L} \frac{n_{x_i,x'_i}}{n_{x_i}} (\lambda_1 - (\lambda_3+1)\Pi_{x_i \neq x'_i} \frac{n_{x_i}}{n_{x'_i}})_+$

$\therefore \int_Z (\lambda_1 d\Pi_X(Z) - (\lambda_3+1)d\Pi_{X'}(Z))_+$

$= |P_X \cap P_{X'}|(\lambda_1 |P_X|^{-1} - (\lambda_3+1)|P_{X'}|^{-1})_+ + \lambda_1 |P_X - P_{X'}||P_X|^{-1}$

$= \Pi_{i=1,x_i \neq x'_i}^{L} \frac{n_{x_i,x'_i}}{n_{x_i}} (\lambda_1 - (\lambda_3+1)\Pi_{x_i \neq x'_i} \frac{n_{x_i}}{n_{x'_i}})_+ + \lambda_1 (1 - \Pi_{i=1,x_i \neq x'_i}^{L} \frac{n_{x_i,x'_i}}{n_{x_i}})$

$where, n_{x_i} = n_{x'_i}$

$= \Pi_{i=1,x_i \neq x'_i}^{L} \frac{n_{x_i,x'_i}}{n_{x_i}} (\lambda_1 - \lambda_3 - 1)_+ + \lambda_1 (1 - \Pi_{i=1,x_i \neq x'_i}^{L} \frac{n_{x_i,x'_i}}{n_{x_i}})$

同理

$\int_Z (\lambda_2 d\Pi_X(Z) - (\lambda_3-1)d\Pi_{X'}(Z))_+$

$= \Pi_{i=1,x_i \neq x'_i}^{L} \frac{n_{x_i,x'_i}}{n_{x_i}} (\lambda_2 - \lambda_3 + 1)_+ + \lambda_2 (1 - \Pi_{i=1,x_i \neq x'_i}^{L} \frac{n_{x_i,x'_i}}{n_{x_i}})$

continue

$\therefore \underset{X' \in S_X}{min}\, g^{RS}(X', c_A) - g^{RS}(X', c_B)$

$= \underset{\lambda_1\ \lambda_2,\lambda_3 \in R}{max} \underset{X' \in S_X}{min} - \int_Z (\lambda_1 d\Pi_X(Z) - (\lambda_3+1)d\Pi_{X'}(Z))_+ - \int_Z (\lambda_2 d\Pi_X(Z) - (\lambda_3-1)d\Pi_{X'}(Z))_+ + \lambda_1 p_A + \lambda_2 p_B - \lambda_3$

$= \underset{\lambda_1,\lambda_2,\lambda_3 \in R}{max} \int_Z (\lambda_1 d\Pi_X(Z) - (\lambda_3+1)d\Pi_{X^*}(Z))_+ - \int_Z (\lambda_2 d\Pi_X(Z) - (\lambda_3-1)d\Pi_{X^*}(Z))_+ + \lambda_1 p_A + \lambda_2 p_B - \lambda_3$

$where, X^* = x_1^*, .., x_L^*$，定义$x_{l_1}, \ldots, x_{l_L}$是按$q_{x_i}$从小到大排序，如果$i \notin [l_1, \ldots, l_R], x_i^* = x_i$，否则，$x_i^* = \underset{x'_i \in S_{x_i}}{argmin}\, n_{x_i,x'_i}/n_x$，证明见原文引理3，$X^*$是最优解。

$= \underset{\lambda_1,\lambda_2,\lambda_3 \in R}{max} - (1-q_X)(\lambda_1 + \lambda_2 - 2\lambda_3)_+ - (\lambda_1 + \lambda_2) q_X + \lambda_1 p_A + \lambda_2 p_B - \lambda_3$

$= \underset{\lambda_3 \in R}{max} - 2\lambda_3 q_X + (\lambda_3+1) p_A + (\lambda_3-1) p_B - \lambda_3$

$= \underset{\lambda_3 \in R}{max} \lambda_3(-2q_X + P_A + P_B - 1) + (p_A - p_B)$

$= p_A - p_B$

$\therefore p_A - p_B > 0 \Rightarrow \underset{X' \in S_X}{min}\, g^{RS}(X', c_A) - g^{RS}(X', c_B) > 0 \Rightarrow$样本X是鲁棒的。

证毕。