中国科学院大学

University of Chinese Academy of Sciences

自然语言处理 命名实体识别 阅读报告

Empower Distantly Supervised Relation Extraction with Collaborative Adversarial Training

2022 年 1 月 1 日

导师：胡玥 教授

姓名：	陈若愚
学号：	202118018629015
学院：	网络空间安全学院
专业：	计算机应用技术

Empower Distantly Supervised Relation Extraction with Collaborative Adversarial Training

Tao Chen[1], Haochen Shi[1], Liyuan Liu[2], Siliang Tang[1*], Jian Shao[1],
Zhigang Chen[3], Yueting Zhuang[1]
1Zhejiang University 2University of Illinois at Urbana Champaign 3iFLYTEK Research
{ttc, hcshi, siliang, jshao, yzhuangg}@zju.edu.cn, llychinalz@gmail.com,
zgchen@iflytek.com

Accepted by *AAAI 2021*

ABSTRACT: 随着远程监督(DS)关系提取(RE)的最新进展，利用多实例学习(MIL)从嘈杂的 DS 中提取高质量监督吸引了相当多的注意力。在这里，我们超越标签噪声，确定 DS-MIL 的关键瓶颈在于其数据利用率低：随着 MIL 对高质量监督的细化，MIL 放弃了大量训练实例，导致数据利用率低和阻碍模型训练有足够的监督。在本文中，我们提出了协同对抗训练以提高数据利用率，在不同级别协调虚拟对抗训练（VAT）和对抗训练（AT）。具体来说，由于 VAT 是无标签的，我们采用实例级 VAT 来回收 MIL 放弃的实例。此外，我们在包级部署 AT，以释放 MIL 获得的高质量监督的全部潜力。我们提出的方法为之前的最先进技术带来了持续的改进（5 个绝对 AUC 分数），这验证了数据利用问题的重要性和我们方法的有效性。

1. Motivation

1.1 Problem

关系提取(RE)的目的是识别特定上下文中实体之间的关系，并为许多下游任务提供必要的支持。由于正则化系统的性能通常受到训练数据量的限制，目前的正则化系统通常采用远程监督(DS)的方法，通过知识库和文本的对齐来获取丰富的训练数据。由于该策略不可避免地会在模型训练中引入标签噪声，如何中和标签噪声一直被视为 DS 的主要问题。

为了从 DS 中提炼出高质量的监督，MIL 只关注少数具有代表性的(注意力得分高的)实例，而放弃了很大比例的低得分实例。如图 1 所示，除了一个 Bag 只包含一个实例（注意分数为 1.0）的情况外，大多数实例的注意分数都很低（0.0~0.2），并且在训练过程中被放弃。

因此作者提出了多实例协同对抗训练（MULTICAST）来提高数据利用率，以解决这个问题，试图补偿它们的数量损失（由 MIL 引起）。

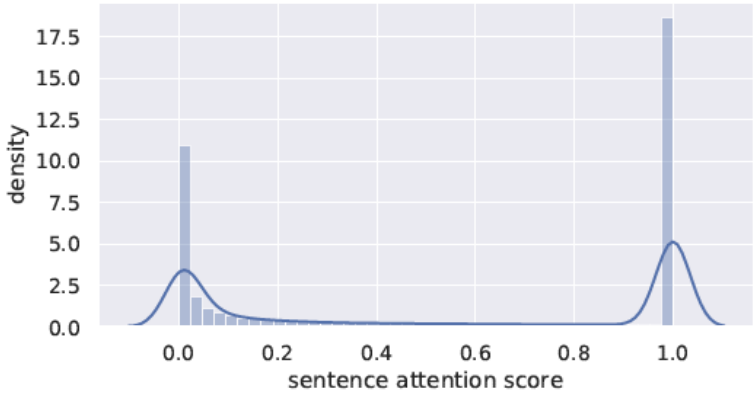

图 1：训练过程中句子注意得分在包内的分布:多数实例得分较低，注意得分较高的实例(不包括 1.0)只占数据的一小部分。

1.2 Contribution

这篇文章的主要贡献如下：

(1) 认为低数据利用率问题是 DS-MIL 的主要瓶颈；

(2) 提出 MULTICAST 以提高数据利用率。它根据 MIL 信号(注意分数)在不同水平上协调 VAT（virtual adversarial training）和 AT；

(3) 在多个细粒度数据集上取得了 SOTA 的结果。

2. Methodology

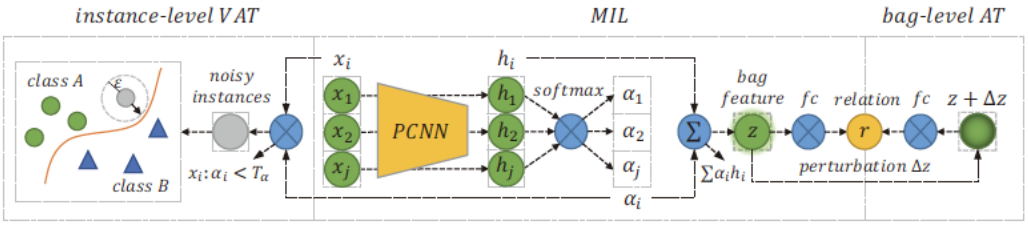

图 2：(a)词袋内实例$x_1, x_2, \cdots, x_j$首先通过分段卷积神经网络对自身进行编码，得到句子级表示$h_1, h_2, \cdots, h_j$。在 MIL 框架的基础上，采用选择性注意的方法在实例上形成较好的词袋级表示$z = \sum_i i\alpha_i h_i$。(b)在包内，选择注意分数α_i较低的嘈杂或不具代表性的实例$\{x_i|\alpha_i < T_\alpha\}$为额外的虚拟对抗训练。(c)词袋外，可靠的袋级表征z通过对抗学习得到进一步增强。

在此论文中，将数据利用率低确定为 DS-MIL 的关键瓶颈。由于 MIL 形成准确的袋子表示来处理标签噪声，因此它放弃了大量的训练实例。通常，MIL 面临标签降噪牺牲数据利用率的困境。提出了协作对抗训练以提高数据利用率。方法(MULTICAST)的图表如图 2 所示，其中包含五个组件：(1)输入表示；(2)句子编码器；(3)基于注意力的 MIL 框架；(4)实例级虚拟对抗训练模块；(5)bag 级对抗训练模块

2.1 Input: Embeddings

这里只是很传统的编码，两部分，一个是关于语义信息的，一个是关于相对信息的，最后将两个表示拼到一起。

对于句子s中的每个词t_i，我们使用词嵌入$w_i \subset \mathbb{R}^{d_w}$来捕获其语义信息。此外，为了以实体感知方式对句子进行编码，利用**相对位置嵌入**来表示句子中的位置信息。相对词t_i的距离d_{i1}，d_{i2}对应于d_{i1}与两个实体e_1和e_2之间的距离，可以通过查找位置嵌入表转移到位置向量$p_{i1}, p_{i2} \subset \mathbb{R}^{d_p}$。这个嵌入表在训练过程中随机初始化和更新。将上述两个嵌入连接(就是 concat)起来，每个词t_i可以获得它的实体感知表示，如$m_i = [w_i; p_{i1}; p_{i2}] \subset \mathbb{R}^d$。因此，实例表示可以构造为$K = [m_1; m_2; \ldots; m_l] \subset \mathbb{R}^{l \times d}$，其中$d = d_w + 2 \cdot d_p$，$l$是句子的最大长度。

2.2 Encoder: Piecewise CNN

这里没什么新颖性，就是传统卷积神经网络，多尺度池化层作为输入。

卷积神经网络通过滑动窗口捕获句子语义。在卷积层，嵌入窗口$X_{t;t+u} = [m_t; m_{t+1}; \ldots; m_{t+u-1}] \subset \mathbb{R}^{u \times d}$与卷积核$\{W_1, \ldots, W_p\} \subset \mathbb{R}^{u \times d}$相互作用，提取句子级特征，其中$u$是内核的宽度，$p$是内核的数量。

紧接着是最大池化层，保留了卷积输出$C \subset \mathbb{R}^{l \times p}$中响应最快的区域。分别将最大池化操作应用于不同的句子片段，这已被证明可以更好地捕获两个实体之间的结构化信息。最终的特征向量$H \subset \mathbb{R}^{3 \times p}$可以通过连接三块的所有池化结果来获得。

2.3 MIL: Multi-Instance Learning

对于由θ参数化的模型，可以将词袋B中每个句子s_i的输入表示$x_i \in X$编码为特征向量$h_i \in H$，然后多实例学习框架考虑包内的所有实例以获得相对准确的表示z，即被定义为：

$$z = \sum_i \alpha_i h_i$$

在权重方面，我们采用了软注意机制，其中i是由基于查询的函数f_i计算的归一化注意分数，该函数测量句子表征h_i和预测关系r的匹配程度：

$$\alpha_i = \frac{e^{f_i}}{\sum_j e^{f_j}}$$

其中$f_i = h_i A q_r$，A是加权对角矩阵，q_r是查询向量，表示关系r（随机初始化）的表示。

然后，在此词袋级表示的基础上，添加一个带有激活函数 softmax 的简单全连接层，将特征向量z映射到条件概率分布上：

$$p(r|Z,\theta)=\frac{e^{o_r}}{\sum_{i=1}^{n_r}e^{o_i}}$$

其中$o=Mz+b$为所有关系类型的得分，n_r为关系总数，M为投影矩阵，b为偏置项。最后，我们使用交叉熵定义 MIL 框架的目标函数如下所示：

$$J(\theta)=-\sum_{i=1}\log p(r_i|z_i,\theta)$$

2.4 IVAT: Instance-Level Virtual Adversarial Training

在 MIL 中，归一化注意力分数i描述了实例x_i对最终表示z的贡献程度。较高的值表示该实例更干净或更具有代表性，而较低的值则表示该实例是嘈杂的（即，其关系标签不可靠）。换句话说，注意力分数是 MIL 中使用的标签质量信号。

将注意力得分高的实例称为X_{clean}，将注意力得分低的实例称为X_{noisy}。MIL 在训练过程中主要关注X_{clean}，而放弃了X_{noisy}。为了提高 MIL 的数据利用率，我们在实例级引入虚拟对抗训练来利用x嘈讯中的实体和上下文信息。

例如$\{x_1,x_2,\cdots,x_i\}$在包B中，我们使用$\{\alpha_1,\alpha_2,\cdots,\alpha_i\}$来参考他们的归一化注意力分数（MIL 部分中选择性注意力的输出）。然后，我们利用超参数T来识别被 MIL 忽略的实例：

$$X_{noisy}=\{x_i|\alpha_i<T_\alpha\}$$

例如$x\in X_{noisy}$，我们把它的条件概率分布输出称为$p(y|x,\theta)$。然后，将其在小扰动$\|d\|\le\epsilon_x$下表示为$x+d$，对应的模型输出为$p(y|x+d,\theta)$。这两个输出被正则化为相似的，即：

$$l_{\text{ivat}}(d,x,\theta):=\text{KL}[p(y|x,\theta)\parallel p(y|x+d,\theta)]$$

KL 是 KL 散度，它度量了两个概率分布的相似性。对于对抗摄动d_{v-adv}，其理想的选择应该是使l_{ivat}最大化的方向。

$$d_{v-adv}:=\arg\underbrace{\max}_{d}\{l_{\text{ivat}}(d,x,\theta);\|d\|_2\le\epsilon_x\}$$

估计 L2 范数下的d_{v-adv}：

$$d_{v-adv}\approx\epsilon_x\frac{g}{\|g\|_2}$$

其中$g=\nabla_r\text{KL}[p(y|x,\theta),p(y|x+d,\theta)]|_{r=\xi d}$，其中$\xi>\mathbf{0}$，$d$是一个随机采样的单位向量。对于神经网络，这种近似可以通过K组反向传播来实现。在这样一个扰动d_{v-adv}下，我们的目标是使模型的局部分布平滑度(LDS)尽可能高，这定义为：

$$\text{LDS}-\text{X}(\theta):=-\sum_{x\in X_{noisy}}l_{\text{ivat}}(d_{v-adv},x,\theta)$$

总结：加入随机噪声，用 KL 散度限制概率分布相似性，保证模型的鲁棒性，同时用对抗训练函数使模型局部的参数分布平滑的解决办法。

2.5 BAT: Bag-Level Adversarial Training

与有噪声的实例不同，我们使用高质量的实例来构造词袋级表示z，它可以更好地匹配关联关系，并允许 MIL 减少标签噪声的影响。在这里，我们利用对抗训练来释放高质量监督的全部潜力。

具体来说，我们将扰动d添加到袋级表示z而不是词嵌入x。与 IVAT 不同，我们使用训练标签代替原始输出来对扰动下的输出进行正则化，即：

$$l_{\text{bat}}(d, z, \theta) := -\log p(r|z+d, \theta)$$

与 IVAT 段的虚拟对抗扰动d_{v-adv}相似，对抗摄动 dadv 在模型输出变化最大的方向上，进一步定义为：

$$d_{adv} := \arg \underbrace{\max}_{d} \{l_{\text{bat}}(d, z, \theta); \|d\|_2 \leq \epsilon_z\}$$

L2 范数下：

$$d_{adv} \approx \epsilon_x \frac{g}{\|g\|_2}$$

其中$g = \nabla_z \log p(r|z, \theta)$，利用神经网络中的反向传播算法可以有效地计算出该算法。在这样的扰动下，我们的最大化目标标记为：

$$\text{LDS} - \text{Z}(\theta) := \sum_{z} l_{\text{bat}}(d_{adv}, z, \theta)$$

总结：和之前的差不多，也是对抗训练，这里针对标签噪声级别进行去噪。

2.6 Objective

总体最大化目标函数为：

$$\mathcal{L} = J(\theta) + \beta_1 \text{LDS} - \text{X}(\theta) + \beta_2 \text{LDS} - \text{Z}(\theta)$$

3. Experiment

3.1 Datasets

DSRE dataset—NYT：使用 2005-2006 年的语料库作为训练集，使用 2007 年的数据作为测试集。其中，训练集由 522,611 个句子、281,270 个实体对、18252 个关系事实组成;测试集由 172,448 个句子、96,678 个实体对、1950 个关系事实组成。对于关系标签，该数据集支持 53 种不同的关系，包括 NA，这意味着实体对之间没有关系。

3.2 Evaluation Metrics

- 绘制了精度-召回率曲线(PR-curve)来显示模型精度和召回率之间的权衡
- 使用了曲线下面积(AUC)度量来评估模型的整体性能
- 精度在 N(P@N)度量（就是 FAR 在某个值时候 P 是多少，P 越高越好）
- 还进行了人为评估。

3.3 Baseline Models

PCNN-ATT

PCNN-ATT+ADV

PCNN-ATT+DSGAN

PCNN-ATT-RA+BAG-ATT

PCNN-ATT+SELF-ATT+[CCL-CT]

PCNN-ATT+DC

3.4 Overall Comparison

表 1：所有比较模型的性能。标记为*的模型引用自原始论文，因为没有公开开源代码。

Method	AUC	P@100	P@200	P@300	P@Mean
PCNN-ATT (Lin et al. 2016)	34.13	73.0	69.0	66.0	69.3
PCNN-ATT+ADV (Wu, Bamman, and Russell 2017)	34.99	80.2	72.1	69.4	73.9
PCNN-ATT-RA+BAG-ATT (Ye and Ling 2019)	35.03	77.0	75.5	72.3	74.9
PCNN-ATT+DSGAN (Qin, Xu, and Wang 2018)	35.19	76.2	70.7	68.4	71.8
PCNN-ATT+SELF-ATT* (Huang and Du 2019)	36.80	81.1	71.6	70.4	74.4
PCNN-ATT+SELF-ATT+CCL-CT* (Huang and Du 2019)	38.10	82.2	79.1	73.1	78.1
PCNN-ATT+MULTICAST (Ours)	**38.78±0.15**	**83.7±1.5**	**79.2±1.0**	**74.2±0.7**	**79.0±0.6**

这个是用 P-R 曲线进行评估的，包围面积越大越好，如下图 3 所示。作者用 P-R 曲线证明了自己的模型效果比其他的 baseline 模型都要好。

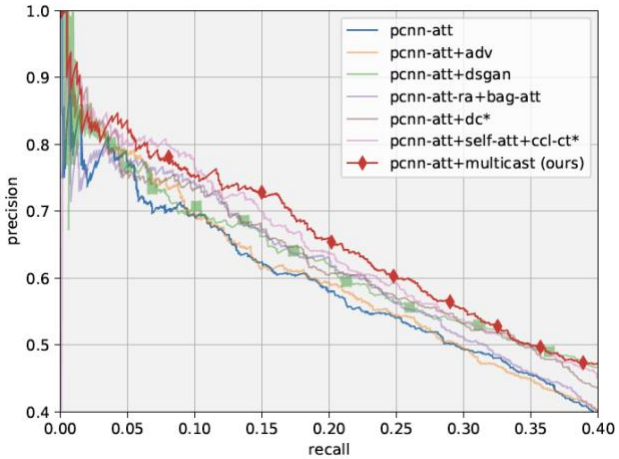

图 3： PR-Curve。带有*的模型直接引用相应论文中绘制的曲线。

在表 1 中，用 AUC 面积和在 N=100，200，400 以及 mean 情况下的 P 值，越大越好，发现模型在这些指标下依然有非常好的效果。

3.5 Controlled Experiment

作者为了验证低分数句子没有被模型使用，我们将这些句子从具有不同阈值的训练集(例如$\alpha_i < 0.1,0.2$)，并使用简化的数据集重新训练 PCNN-ATT 模型和我们提出的模型。评价指标用 AUC，理论上不加会导致 AUC 下降的，如表 2 所示。

表 2：对原始数据集和简化数据集的性能进行建模。

Dataset Size	Method	AUC
522611 (unfiltered)	PCNN-ATT	34.13
	+MULTICAST	38.93
334194(-36%) (filtered @ 0.1)	PCNN-ATT	33.87(-0.7%)
	+MULTICAST	36.50(-6.2%)
310039(-41%) (filtered @ 0.2)	PCNN-ATT	33.70(-1.3%)
	+MULTICAST	36.24(-6.9%)

与预期相符合，因而作者证明了本文提出的方法可以有效地回收 废弃的训练实例，从而实现更好的数据利用率。

3.6 Human Evaluation

与人为评估，我觉得没啥好讨论的，数据如表 3 所示，AUC 越高越好，F1 数值也是越大越好。

表 3：人工标注数据集的建模性能。

Method	AUC	F1
PCNN-ATT	38.91	46.98
PCNN-ATT+DSGAN	43.51(+4.60)	47.49(+0.51)
PCNN-ATT+MULTICAST	**46.03(+7.12)**	**50.29(+3.31)**

3.7 Ablation Study

进一步进行消融研究，以验证我们提出的模块的有效性。消融实验就是展示一个个模块依次叠加的有效性，因为没有理论证明就只能通过实验进行证明了。如表 4 所示，消融实验证明在各个 baseline 上，加入他的方法都最终提高了指标性能，证明各个模块的有效性。

表 4：三种基线模型的消融研究。

Method	AUC	P@100	P@200	P@300	P@Mean
PCNN-ATT (Lin et al. 2016)	34.13	73.0	69.0	66.0	69.3
+BAT	35.10(+0.97)	79.0(+6.0)	77.5(+8.5)	70.7(+4.7)	75.7(+6.4)
+IVAT	37.97(+3.84)	81.2(+8.2)	77.6(+8.6)	73.1(+7.1)	77.3(+8.0)
+IVAT+BAT	**38.93**(+4.80)	**86.2**(+13.2)	**78.6**(+9.6)	**74.1**(+8.1)	**79.6**(+10.3)
PCNN-ATT-RA+BAG-ATT (Ye and Ling 2019)	35.03	77.0	75.5	72.3	74.9
+IVAT*	**38.23**(+3.20)	**87.0**(+10.0)	**82.5**(+7.0)	**75.3**(+3.0)	**81.6**(+6.7)
PCNN-ATT+DSGAN (Qin, Xu, and Wang 2018)	35.19	76.2	70.7	68.4	71.8
+BAT	36.24(+1.05)	79.2(+3.0)	73.1(+2.4)	71.8(+3.4)	74.7(+2.9)
+IVAT	39.21(+4.02)	84.2(+8.0)	77.6(+6.9)	73.4(+5.0)	78.4(+6.6)
+IVAT+BAT	**40.85**(+5.66)	**86.2**(+10.0)	**81.1**(+10.4)	**74.4**(+6.0)	**80.6**(+8.8)

作者同时使用 PR 曲线可视化了 3 中模块依次累加情况下的取消变换，发现曲线稳定上升，用 P-R 曲线进一步证明了方法的有效性。

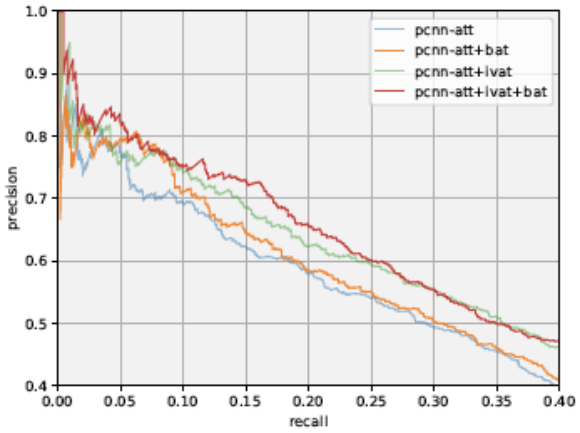

图 4： 模型 IVAT 和 BAT 的 PR-Curve。

3.8 Discussion About AT and VAT

作者讨论了 AT 和 VAT 两种情况下，在所有实例级别和 bag 实例、噪声实例条件下模型的结果，为了证明给定实例等级对模型的影响，结果如表 5 所示。最终结论为：在所有实例中添加 AT 和在包特性中添加 AT 之间的差距很小。直观地说，这两种方法彼此非常相似，而在包级添加 AT 更快(不需要反向传播到嵌入层)。另一方面，向所有更多的实例添加 VAT(这也会变慢)比只向废弃的实例添加 VAT 的性能更差。验证了高质量实例的上下文信息已经被训练算法利用，**不需要对这些实例应用 VAT。**

表 5：讨论不同层次的选择方式。

Method	Level	AUC
PCNN-ATT	-	34.13
PCNN-ATT+AT	all instances	34.99(+0.86)
	bag features	**35.10(+0.97)**
PCNN-ATT+VAT	all instances	37.35(+3.22)
	noisy instances	**37.97(+3.84)**

作者讨论了协同损失的有效性。考虑 AT 与 VAT 之间的合作策略，结果如表 7 所示。结论为：(1)对于实例级的噪声数据，AT 可能会放大错误标签的影响并导致严重的确认偏差问题，这使得模型收敛过快，并且没有学到额外的东西。(2)对于包级的高质量特征，VAT 可能会削弱 MIL 框架提供的原始监督信息，并使模型训练复杂化。对比表 5 和表 7，Instance-Level AT 和 Bag-Level VAT 实际上对模型性能有负面影响。

最后，作者用特例进行了可视化，如表 6 所示。选取一个典型的 bag 来分别说明它们的作用：（1）对于带有 KB 事实的 bag（见表 6），它由三个不同的句子组成。模块 IVAT 关注这些低分(0.19,0.22)的句子。在 IVAT 模块的帮助下，这些句子可以重新考虑它们的概率分布，而无需考虑它们的噪声标签。例如，虽然第三句

（lebron james and his friends used to drive from akron ...）提到了实体对（lebron, james），但实际上并没有表达 live in 的关系。在 IVAT 的帮助下，这个实例成功意识到错误并发现其真正的标签是 NA。(2)同时，BAT 模块专注于由高质量实例形成的准确包特征。在这个包中，最终的表示主要由第 1 句组成（...包括 akron native lebron james），这足以表达当前包标签居住的情况。经过包级别的对抗性增强后，模型更加对具有更高注意力分数的高质量实例有信心（第一个实例的表示接近包级表示）。

图 5 展示了 IVAT 和 BAT 模块的效果图，IVAT 帮助实例 x3 和 y2 找到它们的正确标签。它与 BAT 一起工作，以平滑各自对抗领域的模型输出，这促使模型生成一个更好的分类边界。

表 6： IVAT 和 BAT 模块如何工作的案例研究。

KB Fact: (*lebron james lived_in akron*) Bag Label: */people/person/place_lived*

Sentences	Attention Score		Sentence Label	
	w/o BAT	w/ BAT	w/o IVAT	w/ IVAT
an estimated 40,000 ohio state fans came to town, including the **akron** native **lebron james**, giving this quintessential college town ...	0.59	0.71	lived_in	lived_in
bynum is not another **lebron james**, the high school phenomenon from **akron**, ohio, who was the top draft pick in 2003 and immediately ...	0.19	0.13	NA	borned_in
lebron james and his friends used to drive from **akron**, ohio, fill a few of the empty aquamarine seats in cleveland's downtown ...	0.22	0.16	lived_in	NA

表 7： 讨论不同的协同方式。

Method	AUC
PCNN-ATT	34.13
+Instance-Level AT+Bag-Level VAT	32.34(-1.79)
+Instance-Level AT+Bag-Level AT	34.16(+0.03)
+Instance-Level VAT+Bag-Level VAT	36.36(+2.23)
+Instance-Level VAT+Bag-Level AT	**38.93(+4.80)**

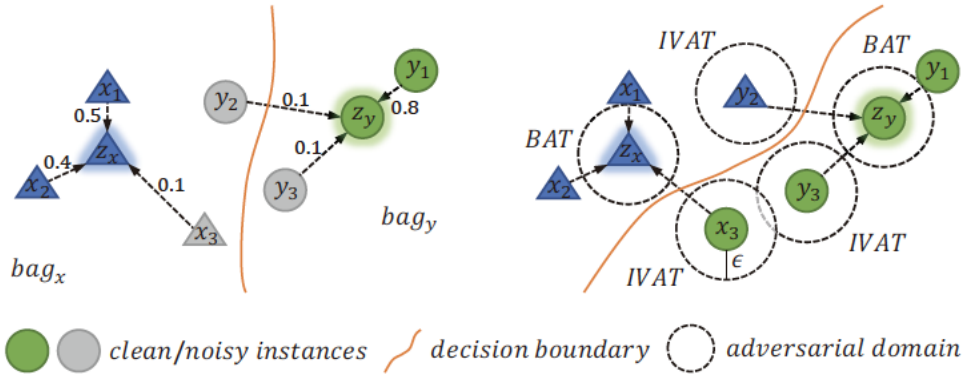

图 5： IVAT 和 BAT 模块效果图。

4. Summary

这篇文章整体来说比较简单，主要是引入了对抗平滑来防止标签中噪声问题，从而使数据因为 MIL 情况导致的数据利用率非常低。作者提出多实例协同对抗训练来缓解，具体来说他们从两个级别，语义级别和标签级别进行平滑处理，通过对

抗学习加入目标函数进行约束，从而尽可能提高了数据的利用率。作者也做了很多消融实验证明其方法的有效性。