北京大学学报(自然科学版) 第 55 卷 第 1 期 2019 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 55, No. 1 (Jan. 2019)
doi: 10.13209/j.0479-8023.2018.064

# 基于对抗学习的讽刺识别研究

张庆林 杜嘉晨 徐睿峰†

哈尔滨工业大学(深圳)计算机科学与技术学院, 深圳 518055; † 通信作者, E-mail: xuruifeng@hit.edu.cn

**摘要** 为了避免现有讽刺识别方法的性能会受训练数据缺乏的影响, 在使用有限标注数据训练的注意力卷积神经网络基础上, 提出一种对抗学习框架, 该框架包含两种互补的对抗学习方法。首先, 提出一种基于对抗样本的学习方法, 应用对抗生成的样本参与模型训练, 以期提高分类器的鲁棒性和泛化能力。进而, 研究基于领域迁移的对抗学习方法, 以期利用跨领域讽刺表达数据, 改善模型在目标领域上的识别性能。在3个讽刺数据集上的实验结果表明, 两种对抗学习方法都能提高讽刺识别的性能, 其中基于领域迁移方法的性能提升更显著; 同时结合两种对抗学习方法能够进一步提高讽刺识别性能。

**关键词** 讽刺识别; 对抗学习; 注意力机制; 卷积神经网络; 对抗样本

# Sarcasm Detection Based on Adversarial Learning

ZHANG Qinglin, DU Jiachen, XU Ruifeng†

School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055;
† Corresponding author, E-mail: xuruifeng@hit.edu.cn

**Abstract** Existing sarcasm detection approaches suffer from lack of sufficient training data. To address this problem, the authors propose an adversarial learning framework built on convolutional neural network (CNN) and attention mechanism, which is trained from limited amounts of labeled data. Two complementary adversarial learning approaches are investigated. First, by training with generated adversarial examples, the authors attempt to enhance the robustness and generalization ability of the classifier. Then, a domain transfer based adversarial learning approach is proposed to leverage cross-domain sarcasm data for improving the performance of sarcasm detection in the target domain. Experimental results on three sarcasm datasets show that both adversarial learning approaches proposed improve the performance of sarcasm detection, but the domain transfer based approach achieves higher performance. Combining the two proposed approaches further improves the performance of sarcasm detection.

**Key words** sarcasm detection; adversarial learning; attention mechanism; convolutional neural network; adversarial examples

近年来, 随着讽刺在微博、论坛等互联网应用中的广泛使用以及文本情感分析问题的深入研究, 越来越多的学者对讽刺识别产生浓厚兴趣。由于用户在使用讽刺表达情感时, 往往出现想表达的情感倾向与字面相反的情况, 所以对讽刺表达的判别会明显影响面向社交媒体的文本情感分析性能。因此, 讽刺识别问题的深入研究对提高文本情感分析系统、问答系统以及会话机器人等自然语言处理应用的性能具有重要意义。

讽刺识别的传统方法主要依靠人工构建特征模板和规则[1−2], 需要依赖领域专家, 且耗费大量的时间和精力, 同时规则系统的可迁移性也比较差。随

国家自然科学基金(U1636103, 61632011)、深圳市基础研究计划(20170307150024907)和深圳市技术攻关项目(JSGG20170817140856618)资助

收稿日期: 2018−06−30; 修回日期: 2018−08−18; 网络出版日期: 2018−08−22

着深度学习模型在众多自然语言处理问题上取得重大突破，有学者将其引入讽刺识别任务中[3–4]。但是，目前讽刺识别领域只有少量公开的人工标注数据或利用弱监督方式自动标注的数据，缺乏大规模、高质量的讽刺标注语料，导致基于机器学习(特别是深度学习)的讽刺识别模型的性能受到一定的限制。

本文提出一种在使用少量标注训练数据的情况下，应用对抗学习框架[5]来提升深度学习模型在讽刺识别任务中性能的方法。首先，在将注意力卷积神经网络[6–8]模型应用于讽刺识别的基础上，采用两种对抗学习方法来提高讽刺识别的性能。其中，基于对抗样本的学习方法[9]在模型训练过程中定向地生成面向识别模型的攻击样本，用于模型训练，以期增强模型的鲁棒性和泛化性能。考虑到基于对抗样本的对抗学习方法只能在单领域数据上生成对抗样本，为了利用更多的跨领域数据，以便提升模型的性能，本文还提出基于领域迁移的对抗学习方法。该方法在目标领域只有少量标注数据的情况下，利用梯度反转层和领域判别器，迁移跨领域的讽刺标注样本，以期提高注意力卷积神经网络模型在目标领域上的性能。最后，将上述两种对抗学习方法相结合，可以进一步提升模型的性能。在IAC的3个讽刺数据集[10]上的实验结果均取得目前已知的最优性能，显示了应用对抗学习在讽刺识别任务上的有效性。

# 1 相关工作

## 1.1 讽刺识别

本文将文本讽刺识别问题视为二分类问题，即给定一条文本，判断文本中是否存在讽刺性表达。Kreuz等[1]基于包含感叹词、标点符号等的词汇特征，构建讽刺自动识别系统。Carvalho等[2]将文本中的表情符号以及特殊字符作为特征来设计讽刺识别算法。近期，也有学者利用深度学习模型搭建讽刺识别系统。Bamman等[3]使用待检测文本的上下文信息，并进一步挖掘社交用户的行为信息，设计基于深度学习的讽刺识别模型。Zhang等[11]使用双向递归神经网络来捕捉目标推特文本的句法和语义信息，同时利用与目标推文相关的历史推文，自动学习特征，进行讽刺识别，并取得较好的性能。Chen等[12]和Gui等[13]从表示学习的角度切入，提高文本情感分类模型的性能。但是，目前大部分基于深度学习的讽刺识别模型均利用小规模人工标注数据训练，性能受到很大限制。也有学者利用网络用户自标注(如hashtag)构建的弱监督数据进行训练，但由于这些数据存在噪音和标签滥用，其文本标签的准确性受到质疑[14]。

## 1.2 对抗训练

联合使用对抗样本和原始样本参与深度学习模型的训练，称为对抗训练。对抗样本指对原始样本增加微小对抗扰动后的样本。对抗样本能够使机器学习算法产生错误的预测，却不会影响人工对样本做出正确分类。Goodfellow等[5]的研究结果表明，对抗训练可以有效地提高神经网络模型防御对抗攻击的能力，从而提高模型的鲁棒性以及泛化性能。Szegedy等[9]首先在计算机视觉领域发现对抗样本的存在，随后Jia等[15]在自然语言处理的相关任务上也发现同样会导致模型性能大幅度下降的对抗样本。Goodfellow等[5]提出的快速梯度法是对抗样本生成中最常用的方法。将基于快速梯度法的对抗训练应用在图像和文本分类领域[5,16–18]，均能提高模型抵制对抗攻击的能力及模型的泛化性能。在文本识别领域，Jia等[15]在模型输入文本的段前或段后等位置随机添加不相关的合法句子或随机字符，生成任务对抗样本，并利用对抗训练来提高阅读理解模型在该任务上的泛化性能。Zhao等[19]利用生成对抗网络来生成图像和文本对抗样本，并将对抗样本用于分析深度学习模型的鲁棒性，增强模型的可解释性。

## 1.3 领域迁移

研究显示，当训练数据和测试数据具有不同的分布时，领域迁移方法可以有效地提高模型性能。Glorot等[20]利用层叠降噪自编码器，学习不同领域间的一致特征表达，提高跨领域文本情感分类性能。Tzeng等[21]通过在卷积神经网络中引入迁移层，在目标损失函数中添加领域混淆损失，训练目标任务模型，在领域迁移的基准任务中取得当时的最优性能。后来，Tzeng等[22]又提出对抗判别式领域迁移模型，解决跨领域手写数字分类问题，提升跨领域手写数字识别的最佳性能。Ganin等[23]利用梯度反转层，最大化领域判别的损失，训练模型学习领域间不变的特征表示。该方法在图像领域迁移任务中均取得当时的最好性能。Gui等[24]通过研究迁移

学习过程中的负迁移问题，提升迁移模型的性能。魏晓聪等[25]提出一种基于Word2Vec的跨领域特征对齐算法，该方法在跨领域情感分类问题上取得较好的性能。

## 2 基于对抗学习的讽刺识别方法

本文将结合注意力机制的卷积神经网络模型作为讽刺识别的基础分类模型。在此基础上，研究基于对抗样本的学习方法进行讽刺识别模型的对抗训练，提高讽刺识别模型的鲁棒性和泛化性能。考虑到基于对抗样本的对抗学习方法只能利用单领域的少量标注数据集来提升模型的效果，进一步研究基于领域迁移的对抗学习方法，使得对抗学习方法能够利用更多的跨领域讽刺数据来提高目标领域的识别性能。最后，本文结合两种对抗学习方法，同时利用对抗样本和跨领域数据集来强化模型的对抗学习过程。

### 2.1 结合注意力机制的卷积神经网络模型

讽刺性文本表达通常由具有共性的短语和表达方式构成。为了保证模型能够捕获这种局部短语和表达方式的共性特征，本文选择卷积神经网络作为基础模型。卷积神经网络主要包括4个部分：输入层、卷积层、池化层和输出层，如图1所示。由于卷积神经网络的最大池化或平均池化的方式会导致文本语义信息的损失，而注意力机制近年来在自然语言处理领域的各类任务中广泛使用，并带来一定的性能提升，因此本文引入注意力机制，将传统卷积神经网络的池化层改为注意力层。通过注意力权值向量，对卷积层输出的特征进行降维和关键信息抽取。

首先，将待识别的文本转化为低维稠密表示向量，组成一个矩阵 $\boldsymbol{S}$，作为模型输入。$\boldsymbol{S}$ 中的第 $i$ 列对应输入文本的第 $i$ 个词 $w_i$ 的词向量 $\boldsymbol{v}_{w_i} \in R^d$。输入矩阵的维数 $N$ 和 $d$ 都是模型的超参数，由人工设定。其中，$d$ 代表词向量的维度，$N$ 代表输入句子的最大长度。超出最大长度 $N$ 的句子会被截去末端的字符，对于句子长度小于 $N$ 的句子，使用零向量 $\boldsymbol{v}_{z_i} \in R^d$ 填充。输入矩阵 $\boldsymbol{S}$ 可表示成如下形式：

$$\boldsymbol{S}=[v_{w_1}, v_{w_2}, v_{w_3}, ..., v_{w_n}, ..., v_{w_{N-1}}, v_{w_N}]。$$

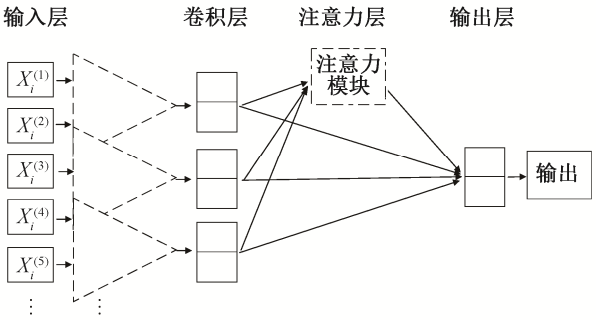


**图 1　结合注意力机制的卷积神经网络模型**

Fig. 1　Convolutional neural network with attention mechanism

在卷积层，不同大小的卷积核在词向量矩阵上平移，进行卷积操作。设某个卷积核 $w \in R^{hd}$，其中 $h$ 是卷积窗口的宽度。输出特征 $c_i \in R$ 的卷积计算过程可形式化地表示为

$$c_i = f(w \cdot s_{i:i+h-1} + b),$$

$f$ 是非线性激活函数，$s_{i:i+h-1}$ 代表 $\boldsymbol{S}$ 中第 $i$ 到 $i+h-1$ 列，$b \in R$ 是偏置项。卷积层的输出为特征 $\boldsymbol{C}$：

$$\boldsymbol{C}=[c_1, c_2, ..., c_{N-h-1}]。$$

注意力机制可以辅助模型捕捉文本中与讽刺分类目标直接相关的关键性文本语义信息。这里，本文结合由Lin等[7]提出的结构化自注意力计算方法，假设卷积层的输出特征矩阵为 $\boldsymbol{C}$，维度为 $R^{n\times m}$。通过注意力计算机制，可以将矩阵 $\boldsymbol{C}$ 转化为固定大小的一维表示向量。注意力计算模块接收特征矩阵 $\boldsymbol{C}$ 作为输入，并输入注意力权值向量 $\boldsymbol{a}$：

$$\boldsymbol{a} = \mathrm{softmax}(\boldsymbol{w}_2 \tanh(\boldsymbol{w}_1 \boldsymbol{C}^{\mathrm{T}})),$$

其中，$\boldsymbol{w}_1$ 是权值矩阵，维度为 $R^{k\times m}$；$\boldsymbol{w}_2$ 为权值向量，维度大小为 $k$。获得注意力权值向量后，将其与输入矩阵相乘，可以快速地获得固定大小的句子或文本表示 $e$。计算公式如下：

$$e = \boldsymbol{a} * \boldsymbol{C}。$$

由于循环神经网络不适用于对文本局部特征建模，所以将未使用结合注意力机制的循环神经网络模型作为分类器。同时，讽刺文本往往长度较大，使用循环神经网络会造成长期遗忘问题，容易导致性能不佳[26]。

### 2.2 基于对抗样本的对抗学习方法

讽刺是一种非常敏感的语言表达方式，细微的语言变化有可能导致模型产生错误的判断。为了提高模型的鲁棒性，本文使用对抗样本和原始样本对讽刺识别模型进行训练(即对抗训练)，使模型可以学习讽刺表达背后真正的语义识别特征。具体地，采用快速梯度法，通过模型的目标损失函数，对输

入数据求梯度，并将其加到相对应的输入维度，从而快速生成对抗样本。框架如图 2 所示。

讽刺分类的神经网络模型输出为 $\text{class}(x) \in \{0, 1\}$。显然，如果分类器模型能够对输入的样本产生高置信度的预测，那么即使对测试样本添加微小扰动，模型也可以做出正确的预测。该过程定义为

$$\tilde{x} = x + \eta \wedge \|\eta\|_\infty < \epsilon \Rightarrow \text{class}(\tilde{x}) = \text{class}(x),$$

这里，$\eta$ 表示添加的噪声扰动，$x$ 是原始样本，$\epsilon$ 是人工设定的超参数，代表添加扰动的最大强度。按照最快梯度法，在每次对抗扰动时，使用一个任意小的正数 $\epsilon$ 来控制添加到原始词向量上扰动的强度，以免改变原始样本的数据分布。在每一步，通过梯度反向传播算法，获得原始词向量最差情况的对抗噪声 $\eta$，从而产生需要的对抗样本。对抗扰动的生成过程可以形式化定义如下:

$$\eta_{\text{adv}} = -\epsilon g / \|g\|_2,\ g = \bigtriangledown_x L(y, \tilde{y}),$$

这里，$g$ 是输入样本 $x$ 的反向传播梯度，$L$ 是模型的目标损失函数。

结合对抗样本的模型损失函数的计算过程可以表示如下:

$$\begin{cases} \text{Loss}_{\text{adv}}(y, \tilde{y}),\ \tilde{y} = F(x + \eta_{\text{adv}}), \\ \text{Loss}_{\text{raw}}(y, \tilde{y}),\ \tilde{y} = F(x), \end{cases}$$
$$\text{Loss} = \alpha \cdot \text{Loss}_{\text{adv}} + (1 - \alpha)\text{Loss}_{\text{raw}},$$

这里，$\alpha$ 是模型的超参数。上式表明，使用对抗样本的对抗训练方法等价于在模型的目标损失函数上增加正则化项，因而对抗训练能够提高模型防御对抗攻击和抵抗过拟合的能力，从而提高模型的泛化性能。特别地，由于讽刺识别任务缺少大型的标注语料，所以在数据层面上，可以借助基于对抗样本的对抗训练方法来提升模型的泛化能力，有效地防止模型出现过拟合。

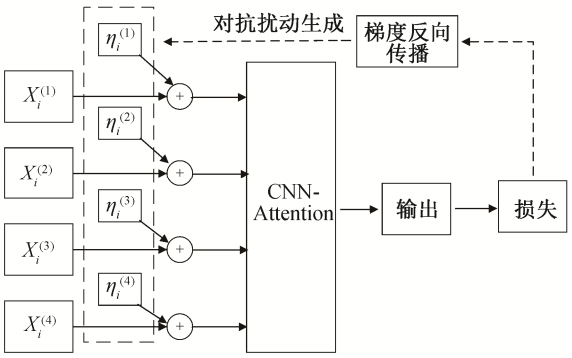


**图 2 基于对抗样本的对抗学习方法**

Fig. 2 Adversarial learning approach based on adversarial examples

## 2.3 基于领域迁移的对抗学习方法

由于单领域的讽刺标注数据存在明显的稀疏，所以结合多领域的讽刺数据集有望进一步提升模型的性能。虽然不同领域的讽刺数据集可能分布差异较大，但可以通过学习领域无关的讽刺语义特征，增强模型的泛化性能。为此，本文研究基于领域迁移的对抗学习方法来训练讽刺识别模型，在包含较多标注数据的源领域训练分类器，对抗迁移至只有少量标注数据的目标领域进行微调和测试。由于训练数据和测试数据具有一定的分布差异，所以普通的训练方法很难在目标领域上取得较好的性能。然而通过领域迁移的对抗学习方法，有望将模型从源领域数据集有效地迁移到目标领域数据集。

领域对抗网络主要通过抽取在目标领域和源领域可迁移的特征表示来降低不同领域数据的分布差异。该方法能够提高深度学习模型在只有少量标注数据的目标领域讽刺识别任务上的性能。该框架主要分为4个部分: 数据输入模块、特征抽取模块(包含注意力计算模块)、讽刺识别模块和领域判别模块，框架如图3所示。

假设，目前拥有较多的源领域标注数据 $D_{\text{s}} = \{(x_i^{\text{s}}, y_i^{\text{s}})\}_{i=1}^{n_{\text{s}}}$ 以及目标领域的无标注数据 $D_{\text{t}} = \{x_j^{\text{t}}\}_{j=1}^{n_{\text{t}}}$，或者少量的标注数据 $D_{\text{t}} = \{(x_j^{\text{t}}, y_j^{\text{t}})\}_{j=1}^{n_{\text{t}}}$(此时 $n_{\text{t}} << n_{\text{s}}$)。理论证明，领域迁移模型在目标领域的性能主要取决于两个因素: 一是目标领域和源领域数据分布的差异性; 二是基于源领域标注数据训练模型的经验误差。本文提出的模型主要目标是将源领域数据和目标领域数据映射到一个共同的特征空间，降低目标领域与源领域数据分布之间的差异性，从而提高只有少量标注数据的目标领域模型的性能。

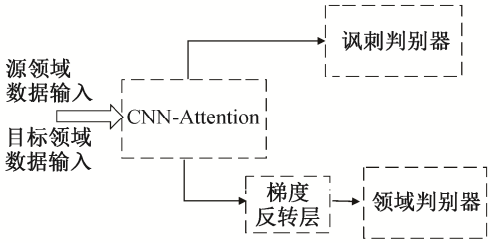


**图 3 基于领域迁移的对抗学习方法**

Fig. 3 Adversarial learning approach based on domain transfer

具体地，在现有的讽刺识别模型中引入一个领域判别器，并在特征抽取模块与领域判别器之间添加梯度反转层。梯度反转层在模型的前向计算和反向传播过程的数学原理可用伪函数$\boldsymbol{R}(x)$形式化地表示为

$$\boldsymbol{R}(x)=x,$$
$$\frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}x}=-\boldsymbol{I},$$

其中，$\boldsymbol{I}$是单位矩阵。梯度反转层在模型的前向计算过程相当于恒等变化，而在模型的误差反向传播学习过程中改变了由领域判别器回传的梯度符号。整个对抗学习策略相当于一个双人博弈游戏，其中一个玩家是领域判决器$G_{\mathrm{d}}$，区分输入的数据来自源领域数据或目标领域数据；另外一个玩家是特征抽取器$G_{\mathrm{f}}$，用来迷惑领域判别器$G_{\mathrm{d}}$，使它无法正确地区分数据来源。

为了抽取领域不变性的特征$f$，特征抽取模块通过最大化领域判别器的损失函数$L_{\mathrm{d}}$来学习参数$\theta_{\mathrm{f}}$。领域判别器通过最小化损失函数$L_{\mathrm{d}}$来调整领域判别器的参数$\theta_{\mathrm{d}}$。整个对抗学习框架的损失函数还包括最小化目标任务(讽刺识别)的损失函数$L_{\mathrm{y}}$。整个领域对抗学习框架的目标代价函数如下:

$$\mathrm{Cost}(\theta_{\mathrm{f}},\theta_{\mathrm{y}},\theta_{\mathrm{d}})=\frac{1}{n_{\mathrm{s}}}\sum_{x_i\in D_{\mathrm{s}}}L_{\mathrm{y}}(G_{\mathrm{y}}(G_{\mathrm{f}}(x_i)),y_i)-\frac{\lambda}{n}\sum_{x_i\in(D_{\mathrm{s}}\cup D_{\mathrm{t}})}L_{\mathrm{d}}(G_{\mathrm{d}}(G_{\mathrm{f}}(x_i)),d_i),$$

其中，$n=n_{\mathrm{s}}+n_{\mathrm{t}}$，$\lambda$是权衡超参数。在模型训练收敛后，参数$\theta_{\mathrm{f}},\theta_{\mathrm{y}}$和$\theta_{\mathrm{d}}$分别收敛于代价函数的一个鞍点，表示如下:

$$(\widetilde{\theta_{\mathrm{f}}},\widetilde{\theta_{\mathrm{y}}})=\arg\min_{\theta_{\mathrm{f}},\theta_{\mathrm{y}}}\mathrm{Cost}(\theta_{\mathrm{f}},\theta_{\mathrm{y}},\theta_{\mathrm{d}}),$$
$$(\widetilde{\theta_{\mathrm{d}}})=\arg\max_{\theta_{\mathrm{d}}}\mathrm{Cost}(\theta_{\mathrm{f}},\theta_{\mathrm{y}},\theta_{\mathrm{d}})。$$

为进一步提高对抗方法的性能，本文研究了同时引入对抗样本和领域迁移的对抗学习框架。在该框架下，对抗扰动的产生只涉及讽刺判别器的目标损失函数$L_{\mathrm{y}}$，与领域判别器的损失函数无关。

# 3 实验结果及分析

## 3.1 实验设置

本文使用3个不同领域的讽刺识别实验数据集，均来自在线辩论语料库(IAC)[26]，分别是Generic数据集、Hyperbole 数据集以及Rhetorical Questions数据集。Hyperbole数据集主要包含夸张讽刺文本；Rhetorical Questions数据集主要包含反讽文本；Generic数据集主要包含普通讽刺文本。3个数据集虽然均为IAC讽刺文本，但是Hyperbole和Rhetirucal Questions数据集为夸张和反问的讽刺手法，与普通讽刺相比，差异较大。同时，从表1的统计数据可以看出，3个数据集的文本统计信息也具有较大的差异。Generic数据集比Hyperbole和Rhetorical Questions含更多的有标注训练数据，因此在基于领域迁移的对抗学习框架中，将Generic数据集设为源领域数据集，而将Hyperbole和Rhetorical Questions数据集分别作为目标领域数据集。

**表1　实验数据集统计情况**

Table 1　Statistics of experimental datasets

| 数据集 | 规模 | 平均句子长度 | 平均句子数量 | 正负样本比例 |
|---|---|---|---|---|
| Generic | 3260 | 51.9 | 2.9 | 1:1 |
| Rhetorical Questions | 850 | 70.3 | 4.5 | 1:1 |
| Hyperbole | 582 | 63.3 | 4.0 | 1:1 |

本文基于3个数据集构造各自的对抗样本，实现基于对抗样本的对抗学习框架。对于每个数据集，随机抽取20%的标注数据作为测试数据，剩余数据作为模型的训练数据和验证数据。以下实验数据均为5次随机实验后的平均性能。

本文使用卷积神经网络作为对抗框架的基模型。卷积神经网络模型输入样本的最大长度$N$设置为300。卷积网络模型使用两种规格的卷积核，宽度分别为3和5。网络的激活函数使用修正线性单元(ReLU)，训练过程中每个批次包含64条样例。预训练的词向量维度设定为300。网络中同时加入dropout层以及L2正则化。模型训练时，将最小化交叉熵损失函数作为模型训练目标，梯度下降法作为模型的优化方法。选择ADAM作为优化器，学习率为$1\times10^{-3}$。

## 3.2 实验结果与分析

为了分析基于对抗样本的对抗训练方法对模型泛化性能提高的效果以及对抗扰动增强模型性能的有效性，本研究增加对原始样本添加随机噪声扰动的被污染样本参与模型训练的对比试验，对比模型如下。

1) CNN-Attention: 不对训练样本做任何修改和数据增强操作的注意力卷积神经网络模型。

2) CNN-Gaussian: 对训练样本添加高斯随机噪声扰动的注意力卷积神经网络模型。

3) CNN-Adv: 对训练样本添加对抗扰动, 生成对抗样本参与模型训练。

对比实验选用高斯噪声, 并将 $\epsilon$ 设置为高斯分布的标准差, 从而控制随机扰动的强度。因此, 第一组实验的对比模型包括普通训练的模型和添加高斯随机扰动训练的模型, 实验结果如表 2 所示。

从表 2 可以看出, 与普通训练模式下的模型CNN-Attention相比, 基于对抗样本的对抗学习模型CNN-Adv的准确率和F1值在 3 个不同的数据集上均有约3个百分点的性能提升, 显示出基于对抗样本的对抗学习方法可以有效地提高模型的泛化性能。相反地, 与CNN-Attention相比, CNN-Gausssian在各数据集的性能均有所下降, 显示添加随机噪声反而降低了模型的泛化性能。这说明, 对抗扰动的添加是提高模型泛化性能的关键因素。在模型训练过程中, 添加对抗扰动有助于定向地降低模型对样本的数值敏感度, 增强模型的泛化性能。相反地, 添加随机扰动并不能起到增强模型泛化性能的作用。

第二组实验评估基于领域迁移的对抗学习方法的性能, 对比模型如下。

1) 基线模型(CNN-Attention): 单独使用源领域数据集(Generic)上训练的注意力卷积神经网络模型。

2) 模型微调(CNN-Finetune): 在源领域训练完成后, 继续使用少量目标领域标注数据进行训练, 微调模型。

3) 基于对抗迁移的模型(CNN-Adversarial_Transfer, CNN-AT): 在基于领域迁移的对抗学习框架下获得的讽刺识别模型。

基于领域迁移的对抗学习方法性能的评估结果如表 3 所示。可以看出, 如果将源领域数据训练的模型直接迁移到CNN-Attention, 由于缺少目标领域数据的训练过程, 其性能与各自领域单独训练的模型相比反而有所下降, 说明目标领域和源领域具有较大的数据分布差异, 导致模型无法在领域间直接迁移。从表 3 中微调迁移模型CNN-Finetune和对抗迁移模型(CNN-AT)的性能比较可以看出, 模型的微调迁移和对抗迁移都能在一定程度上降低跨领域数据集的分布差异。相比而言, 基于领域迁移的对抗学习框架CNN-AT能更有效地增加模型的泛化性能。特别地, 除Hyperbole和Rhetorical Question数据集外, 在Generic数据集上也可以看到模型性能的提升。这从另一个角度说明, 对抗训练能够帮助模型学习到领域无关的讽刺语义特征。

为了进一步提升目标领域讽刺识别的性能, 本文结合领域迁移和对抗样本的对抗学习方法, 第三组实验评估使用该方法后模型的性能。对比模型包括SVM[27]、Deepmoji[28]以及本文的基于对抗样本、

**表 2 基于对抗样本的学习方法实验结果**

Table 2 Experimental results on the learing approach based on adversarial examples

| 模型 | Generic | | Hyperbole | | Rhetorical Questions | |
|---|---|---|---|---|---|---|
| | 准确率 | F1 值 | 准确率 | F1 值 | 准确率 | F1 值 |
| CNN-Attention | 0.731 | 0.732 | 0.629 | 0.635 | 0.681 | 0.681 |
| CNN-Gaussian | 0.664 | 0.671 | 0.585 | 0.586 | 0.620 | 0.621 |
| CNN-Adv | **0.767** | **0.764** | **0.654** | **0.654** | **0.711** | **0.711** |

说明: 粗体数字表示最好结果, 下同。

**表 3 基于领域迁移的对抗学习框架的实验结果**

Table 3 Experimental results on adversarial learning models based on domain transfer

| 模型 | Generic | | Generic → Hyperbole | | Generic → Rhetorical Questions | |
|---|---|---|---|---|---|---|
| | 准确率 | F1 值 | 准确率 | F1 值 | 准确率 | F1 值 |
| CNN-Attention (Gen) | 0.731 | 0.732 | 0.591 | 0.577 | 0.590 | 0.539 |
| CNN-Finetune | 0.731 | 0.732 | 0.644 | 0.644 | 0.718 | 0.718 |
| CNN-AT | **0.765** | **0.765** | **0.704** | **0.704** | **0.732** | **0.723** |

**表 4　结合对抗样本和领域迁移对抗学习框架的实验结果**

Table 4　Experminetal results on the models based on both adversarial examples and domain transfer

| 模型 | Generic | | Hyperbole | | Rhetorical Questions | |
|---|---|---|---|---|---|---|
| | 准确率 | F1 值 | 准确率 | F1 值 | 准确率 | F1 值 |
| SVM(W2V) | — | 0.730 | — | 0.575 | — | 0.680 |
| DeepMoji | — | 0.750 | — | — | — | — |
| RNN-Attention | 0.659 | 0.663 | 0.589 | 0.603 | 0.647 | 0.649 |
| CNN-Attention | 0.731 | 0.732 | 0.629 | 0.635 | 0.681 | 0.681 |
| CNN-Adv | 0.767 | 0.764 | 0.654 | 0.654 | 0.711 | 0.711 |
| CNN-AT | 0.765 | 0.765 | 0.704 | 0.704 | 0.732 | 0.723 |
| CNN-AT-Adv | **0.782** | **0.780** | **0.733** | **0.728** | **0.744** | **0.744** |

基于领域迁移的模型和普通训练模式下的模型。

1) SVM(W2V): 利用预训练好的词向量构建的基于支持向量机的讽刺识别模型。

2) DeepMoji: 利用大规模外部社交情感数据预训练好的深度神经网络讽刺识别模型。

3) CNN-Attention和RNN-Attention: 不对训练样本做任何修改的注意力卷积神经网络模型和递归神经网络。

4) CNN-AT-Adv (CNN-Adversarial_Transfer-Adversarial_Examples): 结合领域迁移和对抗样本的两种对抗学习框架, 共同训练所得的讽刺识别模型。

实验结果如表 4 所示, 可以看出, 结合对抗样本和对抗迁移的方法有效地提高了模型的识别性能, 在3个讽刺识别数据集上取得目前已知最优性能。与现有的两种公开模型对比, 本文的模型在3个讽刺数据集上性能均获得提升。实验结果表明, 数据层面上的基于对抗样本的对抗学习方法和模型层面上的基于领域迁移的对抗学习方法都能有效地提高模型的泛化性能, 缓解深度学习模型在缺少标注数据时的过拟合问题, 从而提高讽刺识别系统的性能。

## 4 结语

本文针对缺少大规模讽刺文本标注数据的情况, 提出两种对抗学习方法, 提升了深度学习模型在讽刺识别上的泛化性能。本文分别研究了基于对抗样本的对抗学习方法和基于领域迁移的对抗学习方法以及两者的结合。本文实现的方法在3个公开的IAC讽刺识别数据集上的实验结果均取得明显的性能提高, 取得目前已知的最优性能, 显示了对抗学习框架在讽刺识别研究中的优越性。然而, 对抗学习框架在训练时仍然存在一些问题, 比如模型训练不稳定, 超参数选择困难等。今后, 将进一步探索对抗学习框架训练时的不稳定问题, 同时更深入地探索对抗样本方法和领域迁移对抗方法在更多自然语言处理问题上的应用。

## 参考文献


[1] Kreuz R J, Caucci G M. Lexical influences on the Perception of Sarcasm // Procceings of the Workshop on Computational Approaches to Figurative Language. New York, 2007: 1–4

[2] Carvalho P, Sarmento L, Silva M, et al. Clues for detecting irony in user-generated contents: oh...!! it's "so easy";-) // Proceedings of the 1st CIKM Workshop on Topic-sentiment Analysis for Mass Opinion. HongKong, 2009: 53–56

[3] Bamman D, Smith A N. Contextualized sarcasm detection on twitter // Proceedings of the International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media. Austin, 2015: 574–577

[4] 刘龙飞, 杨亮, 张绍武, 等. 基于卷积神经网络的微博情感倾向性分析. 中文信息学报, 2015, 29(6): 159–165

[5] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples // Proceedings of International Conference on Learning Representations. San Diego, 2015: 1–10

[6] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention // Proceedings of Conference on Neural Information Processing Systems. Montreal,

2014: 2204–2212

[7] Lin Z, Feng M, Santos C N D, et al. A structured self-attentive sentence embedding [EB/OL]. (2017–03–09)[2018–04–01]. https://arxiv.org/abs/1703.03130

[8] Kim Y. Convolutional neural networks for sentence classification // Proceedings of Empirical Methods in Natural Language Processing. Doha, 2014: 1746–1751

[9] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [EB/OL]. (2014–02–19)[2018–04–01]. https://arxiv.org/abs/1312.6199

[10] Abbott R, Ecker B, Anand P, et al. Internet argument corpus 2.0: an SQL schema for dialogic social media and the corpora to go with it // Proceedings of the Tenth International Conference on Language Resources and Evaluation. Portoro, 2016: 4445–4452

[11] Zhang M, Zhang Y, Fu G. Tweet sarcasm detection using deep neural network // Proceedings of International Conference on Computational Linguistics. Lisbon, 2016: 2449–2460

[12] Chen T, Xu R, He Y, et al. Learning user and product distributed representations using a sequence model for sentiment analysis. IEEE Computational Intelligence Magazine, 2016, 11(3): 34–44

[13] Gui L, Zhou Y, Xu R, et al. Learning representations from heterogeneous network for sentiment classification of product reviews. Knowledge Based Systems, 2017, 124: 34–45

[14] Liebrecht C, Kunneman F, Bosch V A. The perfect solution for detecting sarcasm in tweets #not // Proceedings of the 4th Workshop on Comutational Approaches to Subjectivity, Sentiment and Social Media Analysis. Atlanta, 2013: 29–37

[15] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems // Procceedings of Empirical Methods in Natural Language Processing. Copenhagen, 2017: 2021–2031

[16] Tramer F, Kurakin A, Papernot N, et al. Ensemble adversarial training: attacks and defenses [EB/OL]. (2018–01–30) [2018–04–01]. https://arxiv.org/abs/1705.07204

[17] Miyato T, Dai A M, Goodfellow I. Adversarial training methods for semi-supervised text classification [EB/OL]. (2016–11–07) [2018–04–01]. https://arxiv.org/abs/1605.07725

[18] Wu Y, Bamman D, Russell S. Adversarial training for relation extraction // Proceedings of Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017: 1778–1783

[19] Zhao Z, Dua D, Singh S, et al. Generating natural adversarial examples [EB/OL]. (2018–02–23)[2018–04–01]. https://arxiv.org/abs/1710.11342

[20] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: a deep learning approach // Proceedings of International Conference on Machine Learning. Lille, 2011: 513–520

[21] Tzeng E, Hoffman J, Zhang N, et al. Deep domain confusion: maximizing for domain invariance [EB/OL]. (2014–12–10) [2018–04–01]. https://arxiv.org/abs/1412.3474

[22] Tzeng E, Hoffman J, Saenko K, et al, Adversarial discriminative domain adaptation // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 2962–2971

[23] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation // Proceedings of International Conference on Machine Learning. Lille, 2015: 1180–1189

[24] Gui L, Xu R, Lu Q, et al. Negative transfer detection in transductive transfer learning. International Journal of Machine Learning and Cybernetics, 2018, 9(2): 185–197

[25] 魏晓聪, 林鸿飞. 面向迁移学习的文本特征对齐算法. 计算机工程, 2017, 43(2): 215–219

[26] Pascanu R, Mikolov T, Bengio Y, et al. On the difficulty of training recurrent neural networks // Proceedings of International Conference on Machine Learning. Atlanta, 2013: 1310–1318

[27] Oraby S, Harrison V, Reed L, et al. Creating and characterizing a diverse corpus of sarcasm in dialogue // Proceedings of SIG dial Workshop on Discourse and Dialog. Los Angeles, 2016: 31–41

[28] Felbo B, Mislove A. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm // Proceedings of Empirical Methods in Natural Language Processing. Copenhagen, 2017: 1615–1625