一文读懂文本处理中的对抗训练

Original WenZe、Leo PaperWeekly 2019-06-05

作者丨WenZe、Leo
单位丨追一科技AI Lab研究员

背景与研究意义

深度学习技术的快速发展，大幅提升了众多自然语言处理任务（比如文本分类，机器翻译等）的效果，越来越多的深度学习模型被用于现实生活中。但是深度学习模型本质上的黑箱属性，也为实际应用带来了潜在的风险。

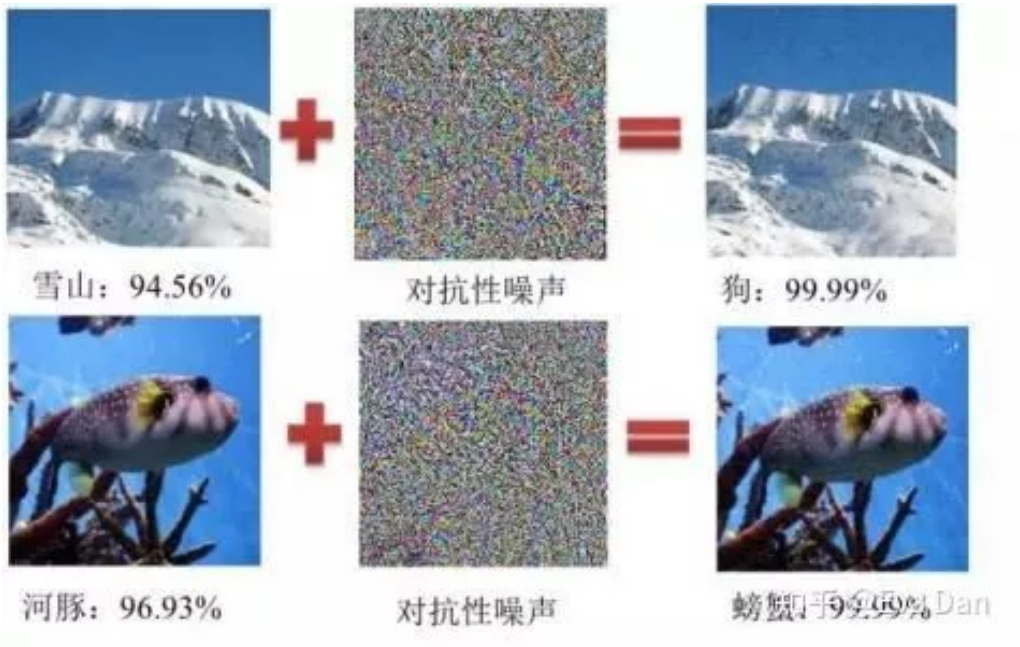
早在 2014 年，*Szegedy et al.* [1] 发现只要对深度学习模型的输入添加一些微小的扰动就能轻易改变模型的预测结果。后续的研究将该种扰动称之为对抗扰动，扰动后的输入称为对抗样本，将输入对抗样本误导模型的这一过程称为对抗攻击。深度学习模型遭遇对抗攻击时所表现出的脆弱性，给实际应用带来了极大的风险。自然语言处理的应用比如文本分类、情感分类、问答系统、推荐系统等也都受到了对抗攻击的威胁 [2]。

在上述背景下，已经有大量的研究集中于提升深度学习模型对于对抗攻击的鲁棒性（也称为对抗防御），其中对抗训练是其中的主要方法之一。**本文对文本处理中的对抗训练进行初步的梳理和总结。**

本文接下来先介绍对抗训练及其相关概念的基本定义，这一部分重点阐述的对抗扰动的基本特征及对应产生扰动的基本方法，由于对抗训练最早开始于图像处理中，其在图像领域中的进展也领先于文本，因此本文第三部分结合其在图像处理领域的研究进展，详细介绍了对抗攻击的基本类型与代表方法，然后简单介绍我们最近在文本分类和鲁棒性上做的实践，最后进行总结。

基本定义与概念

本章节介绍对抗训练 [3] 及相应的基本概念，此段内容主要源于文献 [2]。对抗训练指的是在模型的训练过程中构建对抗样本并将对抗样本和原始样本混合一起训练模型的方法，换句话说就是在模型训练的过程中对模型进行对抗攻击从而提升模型对于对抗攻击的鲁棒性（也称为防御能力）。**可以说不同的对抗攻击方式决定了不同的对抗训练方法。因此文本中的不同对抗攻击方式是本文阐述的重点**，简单的对抗攻击示列如图 1。



▲ 图1. 简单的对抗攻击示列（图片来自https://zhuanlan.zhihu.com/p/37260275）

如图 1 所示（左图为原始样本，中间为添加的对抗扰动，右图为构造的对抗样本），可以看出在加入对抗扰动后原始的图片被误判（雪山变识别成了狗，河豚被识别成了螃蟹），但是人眼并不能够明显发现原图和对抗样本的差异，更不会产生如此离谱的判断。

可以看出对抗攻击指的是在模型原始输入上添加对抗扰动构建对抗样本从而使模型产生错误判断的过程。而在这一过程中，对抗扰动的选择和产生是关键。对抗扰动指的是在模型输入上添加能够误导模型输出的微小变动（图 1 中间部分）。

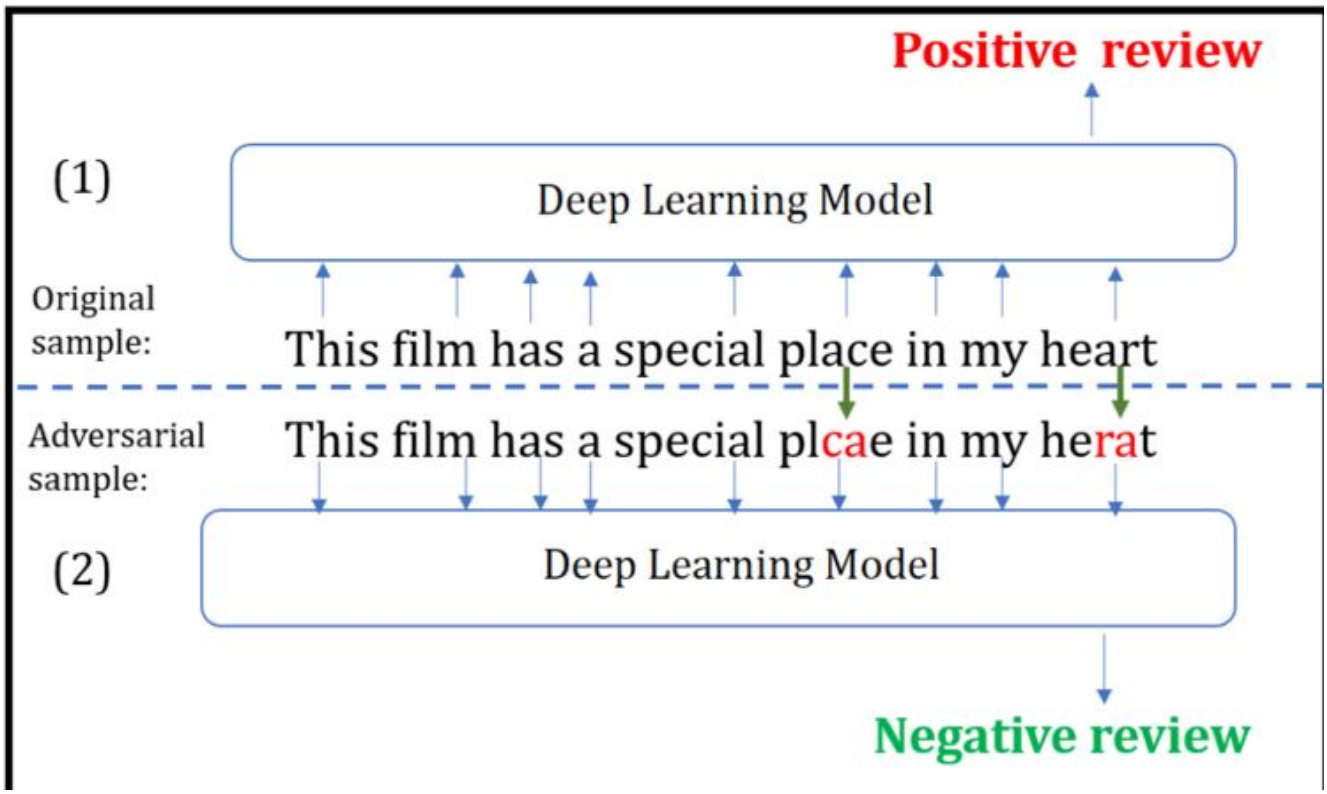
虽然不同的文章对于对抗扰动的定义略有不同，但是一般来说**对抗扰动具有两个特点：**

1. 扰动是微小的甚至是肉眼难以观测到的（图 1 中间部分）；

2. 添加的扰动必须有能力使得模型产生错误的输出（图 1 右侧部分）。

为了满足上述特点，已有大量研究集中于如何产生有效的对抗扰动。不同于图像领域，连续值的扰动直接添加到原始输入矩阵中，在文本处理中添加的扰动可以是离散的也可以是连续的，一般

来说离散扰动指的是直接对输入文本字符进行微小修改（如图 2），连续扰动一般指的是直接在输入文本中的词向量矩阵中添加的扰动 [4]。文本处理中的离散扰动示列如图 2:



▲ 图2. 文本处理中的离散扰动示列（图片来自于[5]）

如图 2 所示，part1 指的是原始的输入文本，part2 指的是对原始数据进行离散扰动后的文本，虽然只有少量字符被修改但是模型产生了完全不同的输出。

在文本处理中，对抗扰动的特征 1 要求添加扰动后产生的对抗样本与原样本在语义上保持一致，即添加的扰动应该尽量不改变原始句子的语义。因此，需要一个测度来衡量扰动前后文本的差异。**下面简单介绍一些在相关文献中已经存在的测度方式**（主要参考自 [2]）。

1. **余弦相似度（Cosine similarity）**是基于词向量的语义相似度计算方法，余弦距离更注重两个向量之间的方向差异。 两个矢量的方向越一致，相似性越大。该方法的局限性在于词向量的维度必须相同。两个给定词向量$\vec{m},\vec{n}$，他们的余弦相似度计算如下：

$$D(\vec{m},\vec{n}) = \frac{\vec{m}.\vec{n}}{\| m \| \cdot \| n \|} = \frac{\sum_{i=1}^{k} m_i \times n_i}{\sqrt{\sum_{i=1}^{k} (m_i)^2} \times \sqrt{\sum_{i=1}^{k} (n_i)^2}}$$

2. **欧式距离（Euclidean Distance）**在文本处理中欧式距离主要应用于连续扰动的情况。给定两个词向量$\vec{m}$与$\vec{n}$，他们的欧式距离计算如下式：

$$D(\vec{m},\vec{n}) = \sqrt{(m_1 - n_1)^2 + \ldots + (m_k - n_k)^2}$$

3. **字移动距离（WMD）**，是 Earth Mover's Distance (EMD) 的变体，它可以通过计算从一个文档到另一个文档嵌入词行进距离来衡量两个文本文档之间的差异。也就是说，WMD 可以量化文本之间的语义相似性。 同时，欧几里德距离也被用于计算 WMD。

4. **Jaccard 相似系数**，给定两个集合 A 和 B，Jaccard 相似系数定义如下式，值越接近与 1，表明两者越相似。在文本处理中，交集就是两者中的相同词，并集就是所有非重复词的集合。

$$J(A,B) = |A \cap B| / |A \cup B|$$

5. **编辑距离（Edit Distance）**是一种通过将字符串转换为另一个字符串来测量最小修改的方法。 它的值越高，两个字符串越不相似。

在上述五种测度中，余弦距离、欧式距离和 WMD 基于词向量计算，主要适用于连续扰动的情况，Jaccard 相似系数，编辑距离可以直接基于文本字符来计算，主要适用于离散扰动的情况。

为了满足对抗扰动的基本特征 2，常用的方法有基于梯度的方法 [3]，直接优化的方法 [1] 等。其中基于梯度的方法一般是计算损失函数对于输入x的梯度dL/dx, 对计算的梯度进行相应变换（*Goodfellow 2014.* [3], *Kurakin A, Goodfellow I 2016.* [6], *Carlini N, Wagner D. 2017.* [7]）从而产生扰动 r。

直接优化的方法，一种是直接修改训练目标以达到攻击目的（比如以分类器遭遇对抗样本时分类出错为目标），通过梯度优化等方式直接找到对抗样本 x_，同时为了满足特征 1，最小化对抗样本与原始样本的距离。

文本中的对抗攻击

对抗攻击可以按照不同的依据分为不同的类别，按对原始模型的访问权限不同可分为：黑盒攻击与白盒攻击。白盒攻击指的是攻击者可以完全访问目标模型，他们可以了解模型的架构，参数和

权重。黑盒攻击指的是攻击者很少或根本没有关于目标模型的知识，他们无法对其进行探测。在这种情况下，攻击者通常训练自己的模型并利用对抗性样本的可转移性来进行攻击。当然，白盒和黑盒攻击都无法改变模型和训练数据。

按攻击目的，可分为目标攻击和非目标攻击，目标攻击指的是生成的对抗样本希望被模型错分到某个特定的类别上。非目标攻击指的是对抗样本只要能让模型分错就行，不论错分到哪个类别都可以。**本文主要阐述对抗攻击中的非目标攻击方法**（接下来的内容主要参考自 [2]）。

不同于图像领域，文本数据的离散特性使得众多研究集中于直接对文本序列进行对抗攻击（对文本序列进行，增删，修改等）。

其中 *Papernot et al.* [8] 首先开始研究文本序列中对抗样本的问题，提出了在递归神经网络（RNN）上产生对抗性输入序列。他们利用计算图展开 [9] 来估算输入序列的正向导数 [10]，即雅可比矩阵。然后对于输入的每个单词，在上述雅可比张量上使用快速梯度符号法（FGSM）[3] 来计算得到扰动。

同时，为了解决修改后的词向量映射问题，他们构建了一个特殊的字典用以选择单词来替换原单词，其中，该替换操作有一个约束，就是替换前后的符号差异要最接近 FGSM 的结果。尽管对抗扰动输入序列可以使 LSTM 模型出现错误，但输入序列的单词是随机选择的，可能存在语法错误。

samanta et al. [11] 引入了三个修改策略，即插入，替换和删除。在尽可能保留输入的语义的前提下，用这些策略生成对抗样本。这些修改策略针对的是那些如果删除后会对分类结果产生很大影响的重要单词。因此，作者利用 FGSM 来验证每一个单词在文本中的贡献，然后以贡献度的递减顺序来定位重要单词。

除了删除之外，插入和替换都需要包括同义词，拼写错误和流派特殊关键词在内的候选池来提供帮助。因此，在实验中，作者为每个单词建立了一个候选池。但是，这样会消耗大量时间，而且事实上，输入文本中很多最重要的单词可能没有候选池。

Gao et al. [12] 提出了算法 DeepWordBug 用以误导 DNN 网络，不同于前面两者，DeepWordBug 适用于黑盒攻击的情境，算法分为两阶段，首先第一阶段是确定哪些重要的 token 要进行改变，第二阶段是产生难以被检测到的扰动。 第一阶段的计算过程如下：

$$CS(x_i) = [F(x_1, \ldots, x_{i-1}, x_i) - F(x_1, x_2, \ldots, x_{i-1})] + \lambda[F(x_i, x_{i+1}, \ldots, x_n) - F(x_{i+1}, \ldots, x_n)]$$

x_i是输入的第 i 个单词，F 是计算置信度的函数。随后，类似如交换，替换，删除和插入等修改策略被应用于重要的 token，这样就得到更好的对抗样本。 同时，为了保持这些对抗样本的可读性，作者使用了编辑距离作为约束测度。

除了上述直接加在文本序列上的对抗扰动方法外，还可以通过在词向量上添加连续扰动的方式进行对抗攻击。*Sato et al.* [4] 直接在 embedding 空间对输入文本上做手脚，用这种方法得到的对抗样本也可以对目标模型进行影响从而导致错误分类。

这种方法的核心思想在于搜索最大化损失函数的方向向量权重，总体参数 W 如下：

$$\alpha_{iAdvT} = \underset{\alpha, \|\alpha\| \leq \epsilon}{argmax}\{\ell(\vec{w} + \sum_{k=1}^{|V|} a_k d_k, \hat{Y}, W)\}$$

其中，$\sum_{k=1}^{|V|} a_k d_k$是从每个输入词向量$\vec{w}$生成的扰动，$\vec{d}$是从 embedding 空间中一个词到另一个的词的方向向量。因为上式非常难以计算，因此作者用下式替代：

$$\alpha_{iAdvT} = \frac{\epsilon g}{\|g\|_2}, g = \nabla_\alpha \ell(\vec{w} + \sum_{k=1}^{|V|} a_k d_k, \hat{Y}, W)$$

iAdvT 的损失函数定义为基于α_{iAdvT}在整个训练数据集 D 上的优化问题，即最小化目标函数：

$$\hat{W} = \frac{1}{|D|} \underset{w}{argmin}\{\sum_{(\hat{X},\hat{Y}) \in D} \ell(\hat{X}, \hat{Y}, W) + \lambda \sum_{(\hat{X},\hat{Y}) \in D} \ell(\hat{X}_{+\gamma(\alpha_{iAdvT}}, \hat{Y}, W)$$

与 *Miyato et al.* [13] 对比，该方法限定了扰动的方向并以此找到预定义词表中的替代词而不是未知词来替换原词。 因此，它通过对抗性训练提高了对抗性例子的可解释性。 同时，作者还利用

了余弦相似性来选择更好的扰动。

相似的，*Gong et al.* [14] 也寻求在 embedding 空间中加入对抗扰动。但是他们的方法是基于梯度的。尽管作者使用 WMD 来衡量干净数据和对抗样本的相似性，但生成结果的可读性较差。

由上述，目前来说文本中的对抗攻击主要分为在文本序列上的离散扰动以及直接作用在词向量矩阵上的连续扰动。最近我们在调研了这些方法之后，也在文本分类上做出了初步的尝试，且取得了不错的效果。

我们的实验

实验采用的文本对抗攻击方式为白盒非目标攻击，构造的扰动是直接基于词向量矩阵的连续扰动。我们分别在文本分类和鲁棒性上进行实验。

文本分类实验数据来自中文公开数据集 ifeng，是 2006-2016 年间凤凰网上的新闻文章，每篇选取前几个段落，数据集有 5 个新闻频道（channel），每个频道（channel）包含的文章数相等，训练集为 80 万，测试集为 5 万。

为了降低训练时间，对原训练集进行了采样。在两个采样训练集中，每个类别的数据量分别为原数据量的 5% 和 10%，未改变数据原有分布，实验数据基本信息如下表：

数据集	训练样本	测试样本	词表大小	类别数目
Ifeng 4000	4000	50000	129742	5
Ifeng 8000	8000	50000	129742	5

▲ 表1. 分类实验数据基本信息

实验方法主要参考自文献 [13][15]，实验结果如下表：

数据集	测试集 size	Baseline	Baseline+ADT
Ifeng 4000	50000	0.7773	**0.7904**
Ifeng 8000	50000	0.7856	**0.7925**

▲ 表2. 文本分类实验对比结果

上表中，Baseline 为现有的分类模型，Baseline+ADT 指的是在现有分类模型的基础上加上对抗训练之后的模型，实验对比指标为分类精度。如上表所示在加入对抗训练后在不同测试集上均有不同程度的提升，其中在小数据集（Ifeng 4000）上提升相对更明显。

此外我们还进行了模型鲁棒性的测试，我们对测试文本进行微小变动（同义词替换，语气词增删等），观察变动前后模型预测结果是否一致。若结果越一致，则鲁棒性越好。即鲁棒性是衡量模型在改动前后预测结果是否一致的评价指标。

对于改动前后的文本，模型预测结果一致则正确数目加一，最终的评价指标为正确数目除以总的测试集数目。实验数据是人为构造的分类数据，最终鲁棒性对比结果如下表：

数据集	Baseline	Baseline+ADT
dataset_1	0.6307	**0.6866**
dataset_2	0.7536	**0.7630**
dataset_3	0.7902	**0.7964**
dataset_4	0.7032	**0.7186**
dataset_5	0.7341	**0.7452**
dataset_6	0.7241	**0.7290**

▲ 表3. 鲁棒性实验对比结果

由上表，在不同训练集合训练出来的模型在加入对抗训练后相对于现有 baseline 在鲁棒性的测试上均有不同程度的提高。

总结

本文介绍了文本中对抗扰动，对抗样本，对抗攻击，对抗训练等基本概念，简单介绍了对抗扰动的特点，不同的对抗扰动产生方法。介绍了对抗攻击的基本类型，结合实列简单介绍了代表性的对抗攻击方法，最后说明了我们最近在文本对抗训练上做的简单实践。

笔者也是近期才开始相关工作，受限于笔者自身的知识水平，难免会有遗漏、说明不当以及错误之处，希望读者能够批评指正。文本中的对抗训练是一个大的知识框架，本文只是简单的给大家

做初步介绍，如果大家感兴趣建议大家阅读最近的一篇综述 [2]，该文章思路清晰，对文本中的对抗攻击和防御进行了系统全面的介绍，本文中的很多内容也都是借鉴上面的，推荐大家阅读。

本文只是简单介绍，并没有针对某一算法进行详细说明，随着自身工作的深入后面有时间会将论文中的具体算法进行实践并写出来分享给大家。

参考文献

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in Proceedings of the International Conference on Learning Representations, 2014.
[2] Wang, W., Wang, L., Tang, B., Wang, R., & Ye, A. (2019). A survey on Adversarial Attacks and Defenses in Text, 1–13. Retrieved from http://arxiv.org/abs/1902.07285
[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in Proceedings of the International Conference on Learning Representations, 2015.
[4] M. Sato, J. Suzuki, H. Shindo, and Y. Matsumoto, “Interpretable adversarial perturbation in input embedding space for text,” in International Joint Conference on Artificial Intelligence (IJCAI), 2018.
[5] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, “Black-box generation of adversarial text sequences to evade deep learning classifiers,” in IEEE Security and Privacy Workshops (SPW). IEEE, 2018.
[6] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.
[7] Carlini N, Wagner D. Towards evaluating the robustness of neural networks//Security and Privacy (SP), 2017 IEEE Symposium on. IEEE, 2017: 39-57.
[8] N. Papernot, P. McDaniel, A. Swami, and R. Harang, “Crafting adversarial input sequences for recurrent neural networks,” in IEEE Military Communications Conference, 2016, p. 4954.
[9] P. J.Werbos, “Generalization of backpropagation with application to a recurrent gas market model,” Neural Networks, vol. 1, no. 4, pp. 339–356, 1988.
[10] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” in IEEE European Symposium on Security and Privacy. IEEE, 2016.
[11] S. Samanta and S. Mehta, “Towards crafting text adversarial samples,” 2017, arXiv preprint arXiv:1707.02812.
[12] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, “Black-box generation of adversarial text sequences to evade deep learning classifiers,” in IEEE Security and Privacy Workshops (SPW). IEEE, 2018.
[13] T. Miyato, A. M. Dai, and I. Goodfellow, “Adversarial training methods for semi-supervised text classification,” in Proceedings of the International Conference on Learning Representations, 2017.
[14] Z. Gong, W. Wang, B. Li, D. Song, and W.-S. Ku, “Ad- versarial texts with gradient methods,” 2018, arXiv preprint arXiv:1801.07175.
[15] Singh Sachan, D., Zaheer, M., & Salakhutdinov, R. (2019). Revisiting LSTM Networks for Semi-Supervised Text Classification via Mixed Objective Function. Retrieved from www.aaai.org.

• END •

点击以下标题查看往期内容推荐：

- ACL 2019 | 基于知识增强的语言表示模型
- 图神经网络综述：模型与应用
- ICLR 2019最佳论文 | 用有序神经元表达层次结构
- F-Principle：初探理解深度学习不能做什么
- 复旦大学邱锡鹏：词法、句法分析研究进展综述
- 基于小样本学习的意图识别冷启动
- 从CNN视角看在自然语言处理上的应用
- 自然语言处理中的语言模型预训练方法

#投 稿 通 道#

让你的论文被更多人看到

如何才能让更多的优质内容以更短路径到达读者群体，缩短读者寻找优质内容的成本呢？ **答案就是：你不认识的人。**

总有一些你不认识的人，知道你想知道的东西。PaperWeekly 或许可以成为一座桥梁，促使不同背景、不同方向的学者和学术灵感相互碰撞，迸发出更多的可能性。

PaperWeekly 鼓励高校实验室或个人，在我们的平台上分享各类优质内容，可以是**最新论文解读**，也可以是**学习心得**或**技术干货**。我们的目的只有一个，让知识真正流动起来。

📝 **来稿标准：**

• 稿件确系个人**原创作品**，来稿需注明作者个人信息（姓名+学校/工作单位+学历/职位+研究方向）

• 如果文章并非首发，请在投稿时提醒并附上所有已发布链接

• PaperWeekly 默认每篇文章都是首发，均会添加“原创”标志

📬 **投稿邮箱：**

• 投稿邮箱：hr@paperweekly.site

• 所有文章配图，请单独在附件中发送

• 请留下即时联系方式（微信或手机），以便我们在编辑发布时和作者沟通

🔍

现在，在**「知乎」**也能找到我们了

进入知乎首页搜索**「PaperWeekly」**

点击**「关注」**订阅我们的专栏吧

关于PaperWeekly

PaperWeekly 是一个推荐、解读、讨论、报道人工智能前沿论文成果的学术平台。如果你研究或从事 AI 领域，欢迎在公众号后台点击**「交流群」**，小助手将把你带入 PaperWeekly 的交流群里。

▽ 点击 | 阅读原文 | 获取最新论文推荐

Read more