人工智能安全

人工智能安全应用 P23

- 网络信息安全应用：主要包括网络安全防护应用、信息内容安全审查应用、数据安全管理应用等
- 社会公共安全应用：主要包括智能安防应用、金融风控应用等

框架安全 P32

Pytorch 相比 TensorFlow 的优势

P35，设计简洁，易于理解，动态计算图

Keras P36 ~ 37

Keras：允许简单而快速的原型设计（由于用户友好，高度模块化，可扩展性）；同时支持卷积神经网络和循环神经网络，以及两者的组合；在 CPU 和 GPU 上无缝运行。

用户友好： Keras 是为人类而不是为机器设计的 API。它把用户体验放在首要和中心位置。将常见用例所需的用户操作数量降至最低，并且在用户错误时提供清晰和可操作的反馈

模块化：模型被理解为由独立的、完全可配置的模块构成的序列或图。神经网络层、损失函数、优化器、初始化方法、激活函数、正则化方法等模块可以以尽可能少的限制组装在一起

易扩展性：新的模块是很容易添加的（作为新的类和函数），现有的模块已经提供了充足的示例

基于Python实现：Keras没有特定格式的单独配置文件。模型定义在Python代码中，这些代码紧凑，易于调试，并且易于扩展

Caffe P38~39

具有表现力的结构：模型及优化是通过配置定义的，而不是使用硬编码的方式。可以在GPU和CPU之间无缝切换，可以用GPU训练，然后部署到集群或移动设备上

代码的可扩展性：Caffe第一年fork超过一千次，有许多有意义的贡献和反馈。由于众多的贡献者，Caffe框架跟踪并实现了当前最新的代码和模型

快速性：快速性使Caffe适合研究实验和工业开发。Caffe在单个的NVIDIA K40 GPU上每天能处理6千万张图片。识别时速度为1ms/张，训练时速度为4ms/张，BLVC相信Caffe具有最快的卷积神经网络实现

框架安全漏洞

1. TensorFlow P41 ~ 43：推理/训练中的拒绝服务攻击，分段错误，在数据流图中插入恶意操作后，不影响模型的正常功能，也就是说模型的使用者从黑盒角度是没有感知的

- Pytorch、Keras 和 Caffe P44 ~ 45：内存泄漏，权重丢失隐患，SQL注入漏洞

环境接触带来的漏洞

1. 第三方基础库漏洞 P47 ~ 50：Numpy 拒绝服务，OpenCV 堆溢出
2. 可移植软件容器漏洞 P51 ~ 52：Kubeflow 挖矿，部署恶意容器

算法安全 P61

鲁棒性安全 P61

- 人工智能算法的优化原理 P62 ~ 63
- 人工智能算法的可解释性 P64 ~ 69

1. 建模前可解释性方法：数据可视化，寻找 ProtoTypes 和 Criticisms（典例与特例）——Prototype 是指能够表示大部分数据的数据实例，Criticism 是指不能由Prototype很好表示出的数据实例
2. 建立本身具备可解释性的模型：一些具有良好可解释性的模型包括决策树模型、线性模型以及贝叶斯实例模型等
3. 使用可解释性方法对模型进行解释：敏感性分析和隐层分析，基于可视化方法的模型解释（结构可视化、训练可视化）

- 人工智能算法的鲁棒性评估 P70

深度学习领域的鲁棒性可以理解为模型对数据变化的容忍度，鲁棒性越高的模型，其识别噪声和对抗样本的准确率越高。鲁棒性差的模型，在受到对抗攻击时，容易给出高可信度的错误结果。鲁棒性差的模型，在更换数据集时（如训练数据更换为测试数据或投入使用时的实际数据），性能往往表现出巨大的差异。

分类维度 P72 ~ 78

- 白盒攻击：攻击者能够获知机器学习所使用的算法，以及算法所使用的参数。攻击者在产生对抗性攻击数据的过程中能够与机器学习的系统有所交互。也就说，攻击者可以将样本送入模型中以获取梯度等信息，然后依据这些信息对输入进行调整
- 黑盒攻击：攻击者对攻击的模型的内部结构，训练参数，防御方法（如果加入了防御手段的话）等等一无所知，只能通过输出输出与模型进行交互。攻击者会使用one-pixel-attack暴力攻击或使用迁移样本进行攻击。这里的迁移样本指的是，对抗样本往往是可迁移特性，即针对机器学习模型A构造的对抗性图像，也会有很大的比例能欺骗机器学习模型B。
- 目标攻击：生成的对抗样本被DNNs误分类为某个指定类别。目标攻击一般发生在多分类问题中。
- 无目标攻击：生成的对抗样本识别结果和原标注无关，即只要攻击成功就好，对抗样本最终属于哪一类不做限制。因为无目标攻击有更多的选择和更大的输出范围，所以比目标攻击更易实现。
- 基于梯度攻击：考虑模型对样本的梯度，根据梯度的方向和大小等对样本进行调整，使损失函数增大
- 基于优化攻击：将输入样本视为可变量，并通过优化算法来最小化或最大化某个特定的目标函数，以找到最优的输入样本，使得模型在该样本上产生误导性的输出结果

安全角度审视机器学习系统 P79

机密性（Confidentiality）——模型隐私，数据隐私；完整性（Integrity）——数据投毒攻击（训练阶段），对抗样本攻击（测试阶段）；可用性（Availability）——系统决策是否准确可靠

投毒攻击 P82 ~ 83

设计攻击样本，混入到训练数据，让人工智能算法失效。

模型学习训练样本的分布，同时假设训练样本和测试样本是同分布的。如果扰乱训练数据的分布，自然会使模型对测试数据给出错误的输出。

P84 与对抗攻击区别

投毒攻击：强调的是通过混入特殊样本的形式，直接对模型进行修改，而不修改测试数据。基于反馈的投毒攻击是指将用户反馈系统武器化来攻击合法用户和内容。一旦攻击者意识到模型利用了用户反馈对模型进行惩罚(penalization)，他们就会基于此为自己谋利。

对抗攻击：攻击者在不改变目标机器学习系统的情况下，通过构造特定输入样本以完成欺骗目标系统的攻击。如基于梯度的攻击方法FGSM等，均属于逃逸攻击。

P88 ~ 89 投毒攻击的防御

为了保障机器学习或深度学习分类模型在受到投毒攻击后的性能，可以在开发模型阶段选取一个干净的数据集，模拟一些投毒策略并据此进行防御。但是，由于可能的攻击空间几乎无限，所以并不能保证防御措施是安全的，即，不能保证一个对已知攻击集有效的防御将不会对新的攻击失效。

对抗样本攻击 P103

对原始数据构造人类难以分辨的扰动，将会引起深度学习算法决策输出的改变，造成人类与深度学习模型认知的差异。近年来对抗样本被证明存在于现实物理世界中，并可能会对多种机器学习系统产生影响。图神经网络(GNN)广泛用于各类场景，然后研究发现图神经网络同样易被实施黑盒对抗样本攻击。各类商用语音识别系统也也易遭受黑盒对抗样本攻击，可以导致定向和非定向攻击效果。

算法局限性 P112

数据难以获取 P112

小数据：数据量太小而无法深度学习。假数据：需要手工生成的数据，有时没有有效性。孤岛数据：A、B数据维度单一，但可以互补，在现实中因为某种原因无法获得两种数据。

数据不完整或偏斜 P114

AI的数据通常包含不完整或偏斜的信息。因为在获取数据时，往往不能获取整个样本空间的数据集。取而代之的是获取一个样本空间的子集。而某些子集的属性并不能代表整个样本空间，因而用这个数据是具有“偏见”的。偏差可以是有意的，也可以是无意的。 数据，算法和选择它们的人员都可能有偏见。偏见可能与种族，性别，年龄，位置或时间有关。

成本局限性 P116

随着训练数据量急剧增长，大模型的训练时间开始以“星期”甚至“月”为单位计量。越来越长的训练时间远远满足不了业务快速迭代的需求。

优秀的人工智能模型背后往往隐藏着巨量的经济开销，主要表现在以下几个方面：数据成本——这一点与数据的局限性相关联，想要获得好的数据就必须付出高昂的成本。开发成本——开发合适的人工智能模型需要一定数量的技术人员，与此对应的财力支出也不可忽略。算力成本——结果显示，算力成本与模型大小成正比，然后对模型调参以提高最终精度的过程中，成本呈爆炸式增长，然而性能收益微乎其微。

算法偏见 P119

AI也同样可能产生偏见，特别是当它向我们人类学习时。**词嵌入**层面就具有偏见。

伦理局限 P122

AI的不可解释性或将长期存在，通过解释某个结果如何得出而实现算法的透明化，在技术上几乎不具有实操性，因而AI的判断也会有错判的情况。如果算法只是帮助我们更好的娱乐、工作，这个问题似乎不那么紧迫。可当算法被用到刑侦、医疗中时，AI被用于给出错误的犯罪者线索、疾病诊断结果时，这是一个必须回答的问题。