# AI安全&算法解决方案鲁棒性的背景

1. 目前AI项目存在的安全风险

1.1 AI安全风险种类划分

1.2 智慧交通中的AI安全风险

2. AI安全技术的扩展性讲解

3. 以数据为中心的鲁棒机器学习策略

3.1 智慧交通与“以数据为中心”策略的结合

3.2 “以数据为中心”策略的具体技术手段

4. 边缘设备软件Aidlux的使用说明

4.1 Aidlux的特点

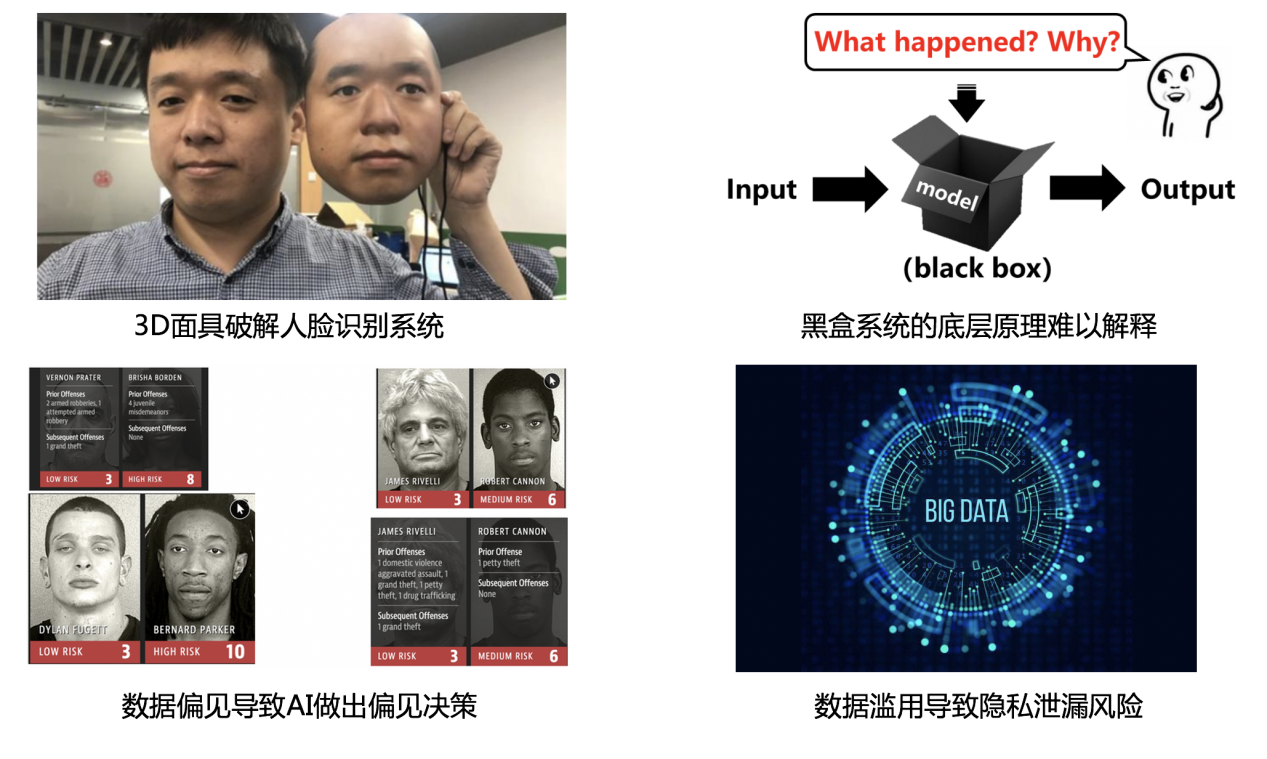
4.2 Aidlux的使用说明和详细操作实例

5. 课程结束

# 1 目前AI项目存在的安全风险

## AI安全风险种类划分

AI项目在不断落地的同时，也伴随着安全风险的日益增加，需要引起工业界的重视。



以深度学习为核心的人工智能存在易受攻击的缺陷，攻击者通过精心制作的3D人脸面具以及相应的合成图片，破解了多个公司的人脸识别系统，说明人工智能的可靠性仍需加强。目前AI算法模型以黑盒为主，具备高度复杂性和不确定性，很多决策原理无法直观理解。这种黑盒特性也大大限制了深度学习理论的进一步发展。AI项目的输出结果受到训练数据的影响，如果数据中存在偏见歧视，AI项目很可能形成偏见的智能决策。比如上图中表示美国芝加哥法院使用的犯罪风险评估系统被证明，对黑人会做出偏见决策。而且AI系统的智能决策也会受到复杂因素影响，导致责任事故主体的界定存在困难。目前最为业界所关注的是自动驾驶与智能机器人等应用出现安全事故时，其责任界定逻辑还需进一步探讨。随着大数据时代的到来，数据滥用可能会导致隐私泄漏的风险，从而影响人们正常的生活社交安全。

## 智慧交通中的AI安全风险

而在智慧交通场景中，面对复杂的交通应用场景，AI算法也很容易受到一些人为的对抗样本攻击，以及复杂自然环境中引入的不确定噪声影响。比如车辆，行人人脸，车牌，交通指示牌等都可能成为受攻击的目标。加上智慧交通AI项目一般都是ToB或者ToG的，客户对项目底层技术的可解释性与数据安全有一定的要求。黑盒系统容易引发不确定性风险，很多交通事件以及交通违规判定仍需专业人员进行二次验证。此外，数据偏见在交通场景中尤为常见。

比如在高速场景中，AI算法往往将卡车，货车等大型车辆作为重点目标，并进行超速检测，安全性分析，交通违规检测等项目功能。相对来说，不够重视小轿车，面包车等小型车辆。可能造成对于小轿车，面包车超速违规的漏检，安全事件的漏判等情况，甚至造成与车辆相级联的相关驾驶员事件功能的偏见决策。

再比如卡口场景，一般来说都是机动车，非机动车等车辆目标，可能会存在对行人相关项目功能的偏见决策与结果输出（行人漏检，机动车虚检，非机动车虚检等）.

在非机动车项目中，外卖小哥每天在外风驰电掣，其流量与行动范围远远大于其他例如自行车，三轮车，家用电动车等非机动车。这些场景中采集的数据必然会存在数据偏见的情况，对模型后续的智能决策会产生影响（比如非机动车驾驶员违规识别偏见决策，非机动车礼让行人偏见决策等）

当然，AI项目一般会持续优化迭代。如果新增的训练数据中一直存在偏见，那么在不断的迭代训练中，算法会进一步受到影响，固化数据中存在的歧视偏见，并形成更加偏见的智能决策。

# AI安全技术的拓展性讲解

前面讲了AI安全的风险，那么如何克服这些风险呢？这里就涉及到AI安全技术了。AI安全技术目前来说，主要包含了安全性，可解释性，公平性以及隐私保护等方面的内容。

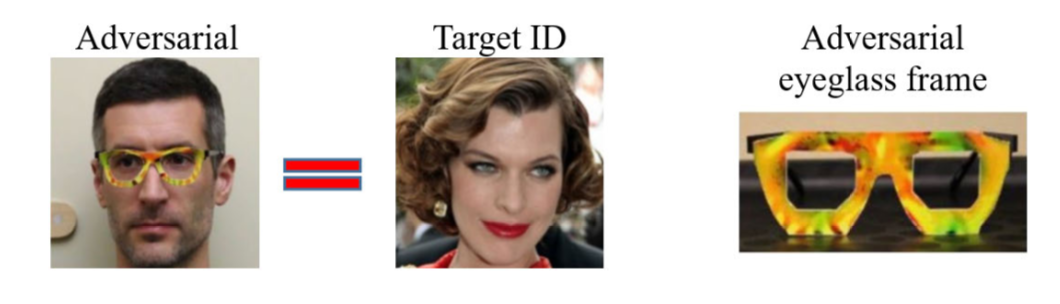


（一）安全性

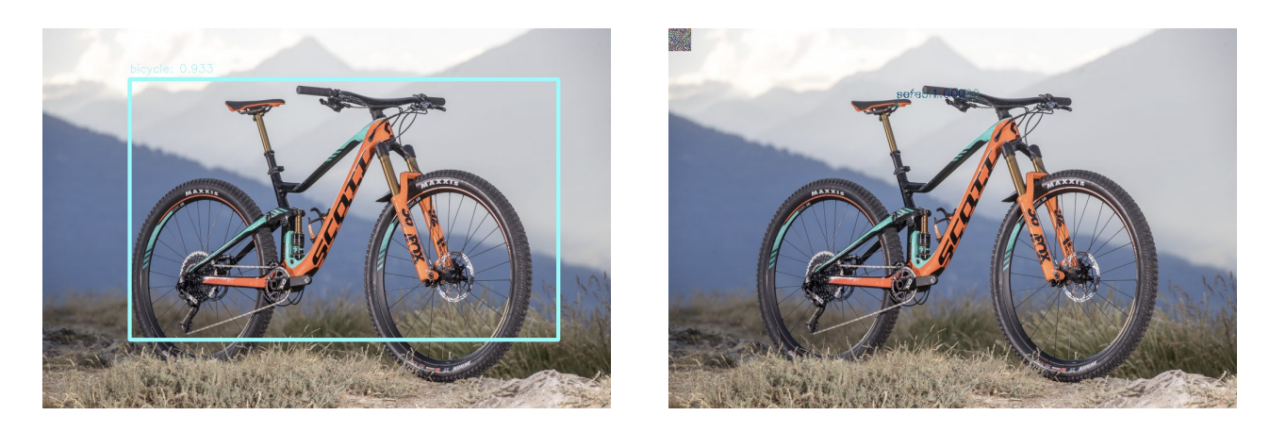
AI系统面临着很多特有的干扰，主要是针对数据和系统的攻击，比如中毒攻击，对抗攻击，后门攻击等。

中毒攻击通过向训练数据投入干扰数据，从而影响AI模型的指标。其多发生在新数据并入训练集的阶段，扰乱数据分布，中毒攻击往往难以察觉。比如在智慧交通中，很多AI项目并不是一次性项目或者demo项目，是会进行不断的优化迭代的。比如卡口场景下的交通事件识别，交通事件的类别会不断丰富，每个交通事件的数据量也会不断增加，不断提升交通事件识别的AI功能的效果。而这个过程中，如果有攻击者将特意制作的干扰样本持续注入新数据流中，在模型侧进行训练时可能会存在中毒攻击的风险。比如行人特征混淆，机动车细分类别混淆，非机动车细分类别混淆等。

对抗攻击主要通过制造对抗样本，让AI模型产生错判，从而导致严重的后果。这在第一节课中已经讲过，大家可以在第一节课的智慧交通+AI安全功能小节中温习一下。这里Rocky展示一些新的例子带着大家再过一遍（后续课程中将重点展开）：



比如使用对抗眼镜攻击让人脸识别系统失效，攻击者戴上一个精心制作的对抗眼镜，使得人脸系统对人脸ID出现了误判，对人脸识别系统整体性能产生较大影响。而当智慧交通中的摄像头设备，接收此类对抗样本并将其作为后续算法功能的输入，可能会导致包含人脸识别的算法解决方案出现不确定风险。



再比如对交通工具进行对抗攻击，在图片左上角添加一个对抗噪声，其并不会破坏交通工具本身的主要特征，但是却能使得使得检测模型对自行车产生了漏检，而重要交通工具的漏检可能引发严重的交通连锁后果。而智慧交通中的摄像头设备，也会接收此类对抗样本并将其作为后续算法功能的输入，从而导致交通算法解决方案中的某一环失效。

我们再来了解一下后门攻击，后门攻击具有很强的隐蔽性，主要是通过对AI系统植入后门，进而对AI系统进行操控，在一些安全要求很高的任务比如自动驾驶，交通安全预警等，可能会带来很严重的后果。在通常情况下，植入后门的模型表现得与干净模型一样正常，因此仅通过验证测试样本的测试准确性来区分植入后门的模型和干净模型是不可能的。当秘密触发器Trigger（只有攻击者知道）出现在输入中时，后门模型就会被错误引导去执行攻击者的子任务，比如分类任务中将输出分类到攻击者指定的目标类别。

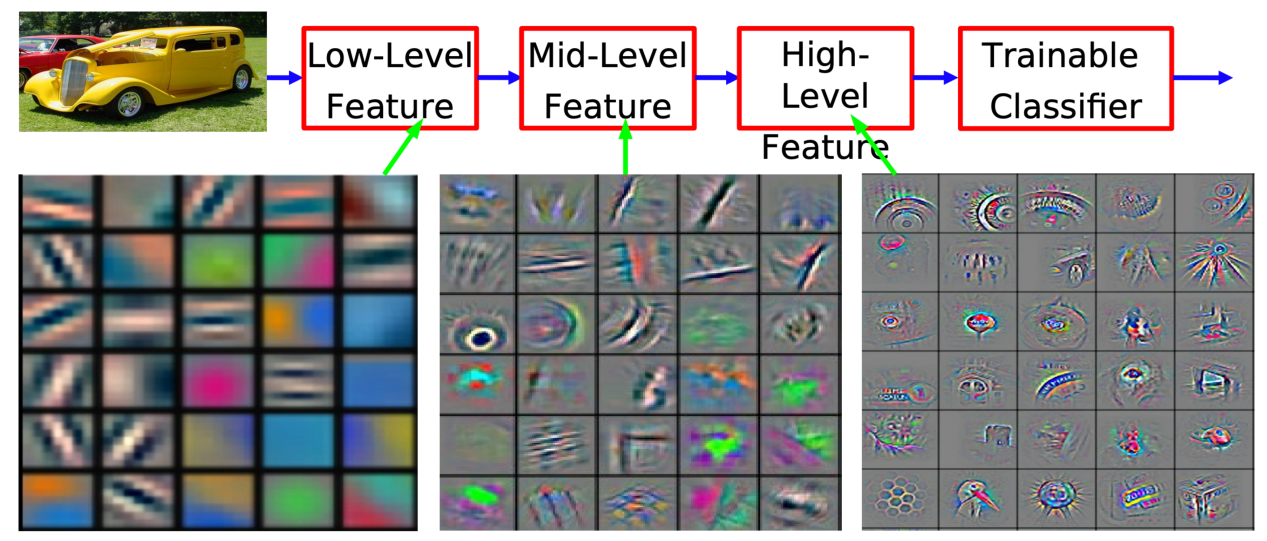
例如，自动驾驶系统可能会被劫持，通过在停车标志上粘贴便利贴，将停车标志归类为限速标志，这可能会导致交通事故。



针对这些攻击，很多工作提出了各种不同的方法，比如设计异常数据检测方法来检测并清除中毒样本，对抗样本，后门样本等攻击性样本；再比如，设计对抗攻击检测技术来避免在线的对抗攻击；或者使用对抗训练，预处理后处理等技术防御对抗样本；通过模型结构鲁棒性设计，模型剪枝，门后检测等技术抵御后门攻击。

（二）可解释性

针对AI模型可解释性弱的问题，一种优化方法是建立可视化机制和解释模型的中间状态来缓解；通过Grad-CAM（Gradient-weighted Class Activation Mapping）等技术分析AI模型的注意力机制；



并通过建立end-to-end的模型训练管理流程来提升AI系统的可复现性，比如Google的TesnorBoard等模型训练管理工具。当然的，很多公司也会自建包含数据+模型+环境+训练+测试的一体化AI训练平台。

（三）隐私保护

针对隐私保护的问题，最常用的是基于差分隐私和基于联邦学习的隐私保护策略，甚至将联邦学习和差分隐私相结合，来构建隐私保护能力更强的AI系统。

差分隐私主要通过随机算法对数据加上一些扰动，避免攻击者分辨出任何具体的隐私或者敏感信息。联邦学习的目标是在保证数据隐私安全及合法合规的基础上，实现数据本地保存，模型线上联合训练，提升AI模型的效果，并一定程度上解决了数据孤岛问题。为了大家更好的理解联邦学习的概念，Rocky在这给大家推荐一篇关于联邦学习概念的漫画图解，非常生动有趣，大家可以参考看看：<https://rockyding.blog.csdn.net/article/details/103162649>

（四）公平性

针对公平性问题，其主要受不同类别在数据中占比不均衡，并且AI模型在这样的数据集上训练而导致偏见决策，所以如何构建完整的异构数据集就成为了重中之重，其可以将数据集中的固有偏见最小化。此外，对数据集进行定期检查与维护，从而保证数据的高质量。并设计删除敏感信息或者不同采样方法等预处理手段，降低数据集中存在的偏差。

Rocky将上面讲的AI安全技术进行了汇总，如下图所示。目前为止，AI安全技术中对抗攻防与联邦学习在业界有比较大的进展。



随着AI项目不断落地，AI人工智能不断对社会生活产生影响，实际应用中的安全性成为了各国和各公司除了性能指标外的首要关切，各国也不断发布相关政策与法规，各公司也不断提出内部的AI安全性要求。因此，本课程后续也将主要讨论AI安全技术中的安全性部分，并从对抗攻击与防御入手，在后续的章节中将深入讲解，并将过程直观的展示。

# 以数据为中心的鲁棒机器学习策略

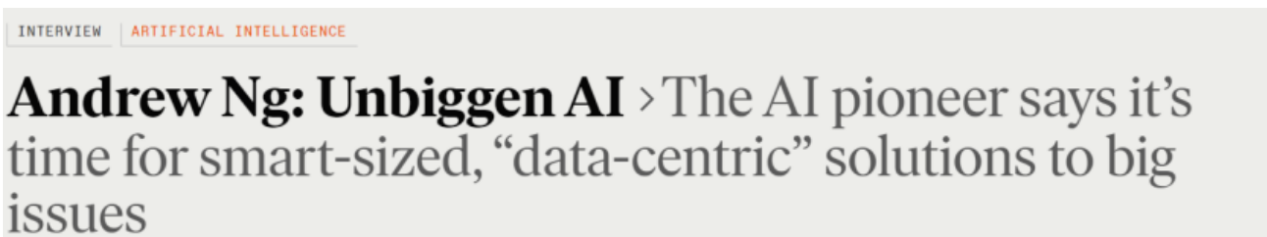
## 智慧交通与“以数据为中心”策略的结合

讲完AI项目目前存在的安全风险以及AI安全技术的应用，我们可以了解到提升AI项目安全的途径，有针对数据流的，针对预处理的，针对模型本身的，以及针对后处理的。当然，从AI算法的角度来说，数据是非常关键的因素。本小节也将从数据部分入手，结合工业界中普遍存在的数据获取难，成本高的问题，着重讲讲用“以数据为中心”的鲁棒机器学习策略来提升模型的准确性与安全性。在智慧交通场景中，主要存在两个大的增长点：

一是通过大场景解决方案，包括之前讲到过的交通卡口，十字路口，出入口以及停车场场景的车辆检测，行人检测，车牌识别，非机动车检测，交通事件等算法功能；

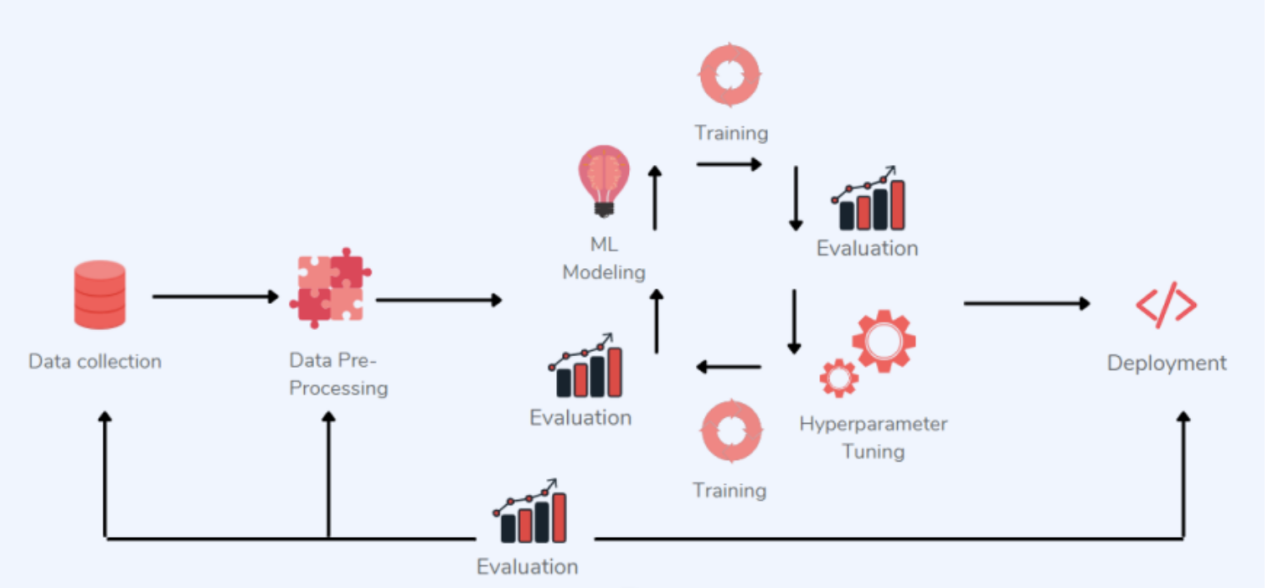
二是交通场景细分领域，包括使用车辆姿态识别，车辆方向识别，道路抛洒物检测，高速场景危险事件告警，交通场景中的稀缺事件判断等算法功能。

这两者分别对应了大数据与稀缺数据。在大场景解决方案落地后，能保证持续的运营维护和现金流。在此基础上，逐步挖掘一些细分场景和长尾场景，加深算法解决方案的护城河。而“以数据为中心“旨在将关注点放在构建AI系统所需的数据上。对许多AI应用场景来说，神经网络架构已经相当成熟，整个训练流程的上下游（模型搭建，模型部署）也已经非常高效，模型及其训练过程不会成为业务场景中的难点。刚才讲的第二种交通场景细分领域中，随着神经网络架构的成熟以及理论难以再有大刀阔斧的革新，对于许多实际应用来说，瓶颈将会存在于“如何获取、开发所需要的数据”。在这些细分领域中，很多情况下高价值数据采集获取的成本非常高，难度非常大。这时就可以使用“以数据为中心”的思想，在给定固定模型的情况下改进稀缺数据集，将重点从“大数据转移到高质量数据”，从而提升模型的鲁棒性与准确性。只要一定量高质量的图片，仍然可以产生非常有价值的东西，例如高速场景危险事件告警。



## “以数据为中心”策略的具体技术手段

在只拥有一定量的高质量数据的情况下，如何通过科学的方法向神经网络解释你想让它学习什么，会成为搭建AI系统非常重要的一环。



具体的一些策略包括数据标注纠偏纠错，对数据子集进行优化，合成特定数据，特定数据增强，标签细化，多模态融合等。比如通过高阶数据增强技术（GAN等生成模型）合成一些交通场景中的长尾数据，包括特种车辆，稀有车牌，稀有车型等数据来支持细分领域的算法功能优化。在数据量级不大的情况下，不断优化数据，生成高质量数据，提供能支持模型优化bad case与算法功能实现的数据，就是“以数据为中心”的核心思想。



总的来说，过去十年人工智能最大的转变是AI理论持续研究，而接下来的十年，相信以数据为中心会帮助AI算法持续在工业界落地。在后面一节的课程里，我们也将使用“以数据为中心”的策略优化我们的车辆检测模型。

# 边缘设备软件Aidlux的使用说明

## Aidlux的特点

目前为止，我们已经了解了AI安全相关的知识以及“以数据为中心”的机器学习策略，加上第一节课学习的智慧交通和CV项目开发落地的理论知识，我们已经有了足够的知识储备，可以开始智慧交通与AI安全项目的学习了。不过在此之前，我们需要先熟悉一下Aidlux边缘设备软件，因为后续我们将在Aidlux上完成很多项目功能。

Aidlux主打的是基于ARM架构的跨生态(Android/鸿蒙+Linux)一站式AIOT应用开发平台。用比较简单的方式理解，我们平时编写训练模型，测试模型的时候，常用的是 Linux/window系统。而实际应用到现场的时候，通常会以几种形态：GPU服务器、嵌入式设备(比如Android手机、人脸识别闸机，摄像头等)、边缘设备。GPU服务器我们好理解，而Android 嵌入式设备的底层芯片，通常是ARM架构。而Linux底层也是ARM架构，并且Android又是基于Linux内核开发的操作系统，两者可以共享Linux内核。因此就产生了从底层开发一套应用系统的方式，在此基础上同时带来原生Android和原生Linux使用体验。



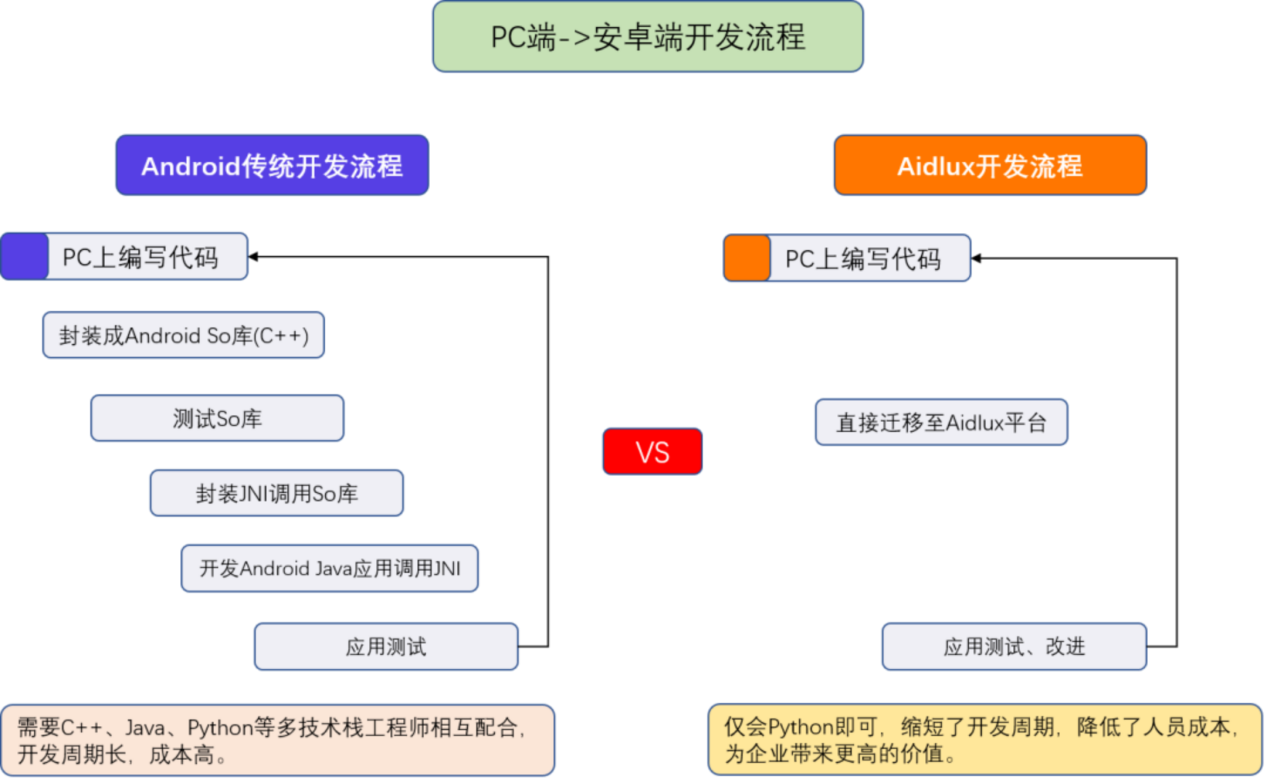
因此基于ARM芯片，开发了Aidlux平台，可以在安卓手机上直接下载Aidlux使用。同时基于 ARM芯片，比如高通骁龙的855芯片和865芯片，也开发了AidBox边缘设备，提供7T OPS和15TOPS算力，可以直接在设备上使用。



而使用这些设备平台开发的时候，和在Linux上开发都是通用的，即Linux上开发的Python代 码，可以在安卓手机、边缘设备上转换后无缝使用。

那么为什么可以无缝使用呢？

常规的方式，应用在手机Android时，需要将PC上编写的代码，封装成Android SO库（C++）。经过测试后，封装JNI调用SO库，最终在Android上使用Java调用JNI，最终再进行测试发布。因此我们可以看到，这样的流程需要一系列的工作人员参与，比如C++、Java、Python的工程师。Aidlux将其中的整个开发流程，全部打通，通过Aidlux平台，可以将PC端编写的代码，快速应用到Android系统上。



那么有了Android和Linux双系统开发的基础，就可以做很多的事情了。而AI算法应用，就是比较典型的一种。不过这里就涉及到，芯片对于AI算法的优化加速的能力。Aidlux内部一方面内置了多种深度学习框架，便于快速开发。另外对于多种算子进行了优化加速，很多算法的性能，也都能达到实时使用。



## Aidlux的使用说明和详细操作实例

因为后面的课程中，需要使用Aidlux，因此大家可以先提前下载Aidlux软件。关于Aidlux的使用说明和详细操作实例，江大白老师已经在上一次训练营的第二节课中讲过。大家可以看一下大白老师第二节课的课程内容：<https://docs.qq.com/doc/DWEdSV2ZwdHVEQWJv>

大白老师的文章框架主要是：

1 AI项目开发及Aidlux的特点

1.1 跨平台应用的系统

1.2 安卓&Linux移植开发流程

1.3 Aidlux算法优化

2 手机版本Aidlux软件安装

3 电脑版Aidlux投影测试

4 Aidlux系统AI案例测试

5 Aidlux软件设置默认后台运行

6 Aidlux&VScode编程调试方式

7 安装本地版的Python和Opencv

8 PC端读取图片&视频操作

9 PC端远程调试Aidlux

10 PC端调试Aidlux读取图片&视频

如果上一次训练营，大家没有参加过的话，可以从第2部分开始看起。如果想要测试里面的读取图片，和读取适配操作，可以在在aidlux的公众号后台回复，lesson2，下载相关文档与测试代码。

Jarvis:

# 5 本节课结束

本节课的内容到这里就结束了，第二节主要是先从AI项目目前存在的安全风险&AI安全技术方面和大家一起了解行业现状。然后一起学习了“以数据为中心”的鲁棒机器学习策略和边缘设备软件Aidlux的相关知识，大家也可以慢慢温习下本节课的内容。第三节课开始就将进行算法模型的训练测试以及部署，将开始与Aidlux平台深度交互，大家可以尽快熟悉Aidlux平台。