CAAI AIDL 演讲实录丨金连文：“文字检测与识别：现状及展望”

中国人工智能学会 9月3日

> 8月31日-9月1日，由中国人工智能学会主办，华中科技大学电子信息与通信学院承办，主题为《计算机视觉应用技术》的AI前沿讲习班第七期在华中科技大学成功举办。
> 在讲习班上，华南理工大学二级教授、博士生导师金连文发表了主题为《文字检测与识别：现状及展望》的精彩演讲。

金连文
华南理工大学二级教授，博士生导师

以下是金连文的演讲实录：

金连文：谢谢许老师的介绍，也谢谢组委会的组织及邀请。今天很高兴来这里做一个关于场景文字检测识别最新进展的报告，来到华中科技大学做这个报告压力很大，因为大家知道华中科技大学白老师团队在场景文字检测和识别领域做得非常好，比我们好多了，所以来这里感觉有一点班门弄斧，心中惶惶然…，所以我尽量讲一点有差异化的内容。

我今天报告主题是文字检测和识别的一些新思考，特别是过去三年以来一些新的进展及发展趋势，并介绍我们实验室做的几个相关工作。

大家知道文字是我们信息交流的最重要的一个媒介，我们生活当中文字是无处不在，可以说离开了文字我们衣食住行各方面都会很不方便。

有这么一句话叫一图胜千言，但其实图中文字信息更重要，离开了文字我们有时很难去理解图像真正的含义。比如这里有两张图，左边是我在法国一个酒店里面拍的，没有文字说明大家很难猜它到底是什么意思，其实这个图是酒店男洗手间的标识！右边的图是一个很萌的小和尚，如果没有文字说明也不知道他在干什么，上面有文字以后这个图就变得有意义了，它引用了金刚经里面的一段文字“一切法无我，得成于忍”，告诉我们做人做事要忍让的道理。所以说文字的信息是很重要的。

这个图大家晃眼一看像药品，但它实际上是什么呢？实际上是葡萄干，光靠图像识别，没有上面的文字帮助，人很难理解这个图像。所以我有一个观点，如果一张图上面有文字，在绝大多数的情况下，图中的文字信息是最重要的，所以文字识别在这个意义上来讲它的重要性恐怕要大于图像识别的重要性。

另外一个方面，文字从整个文化的角度来讲也是非常重要的，人类的文明离不开文字，文字是我们学习知识、传播信息、记录思想很重要的载体，没有文字人类的文明无从谈起。所以我们看到文字识别也是目前最有应用前景的AI人工技术，它在很多方面都有非常广阔的应用前景，比如说图像搜索（文字辅助理解图像信息）、自动驾驶（路标交通标识等识别）、医疗健康、教育产业（比如现在自动阅卷和作业的批改），包括金融、保险等等行业都有很大的应用前景。国内现在很多知名的IT公司，比

如像百度、阿里巴巴、腾讯、科大讯飞等公司，他们都要专门有OCR的团队；很多新的人工智能创业的公司，比如说商汤、旷世、云从、合合信息等，OCR都是他们重要的落地应用技术之一。

文字识别领域的应用场景非常广阔，据我了解，从金融行业、电子商务到消费电子等许多领域都有非常强的需求，例如包括平安银行、招商集团、京东、美团、搜狗、华为、联想、VIVO、金山、海康威视甚至到美的集团等等许多企业都有不同的OCR的技术需求。

在讲今天的报告之前有三个名字先解释一下，分别是OCR、DAR、STR，OCR就是光学字符识别（Optical Character Recognition）的英语术语首字母的缩写，表示通过光学数字化设备（比如扫描仪、数码相机等）拍摄的图像，对图像中的文字进行识别。在文档图像处理领域还有一个更大的概念叫文档图像分析和识别（Document Analysis & Recognition），我们通常把它叫DAR，这是一个更广泛的概念。另外一个比较细的领域就是场景文字识别STR（Scene Text Recognition），主要是针对自然场景当中以手机为主的拍摄图片从中去检测和识别文字。这是三个不同的概念。大家通常听得比较多的是OCR这个词，场景文字识别实际上是OCR的一种典型分支。

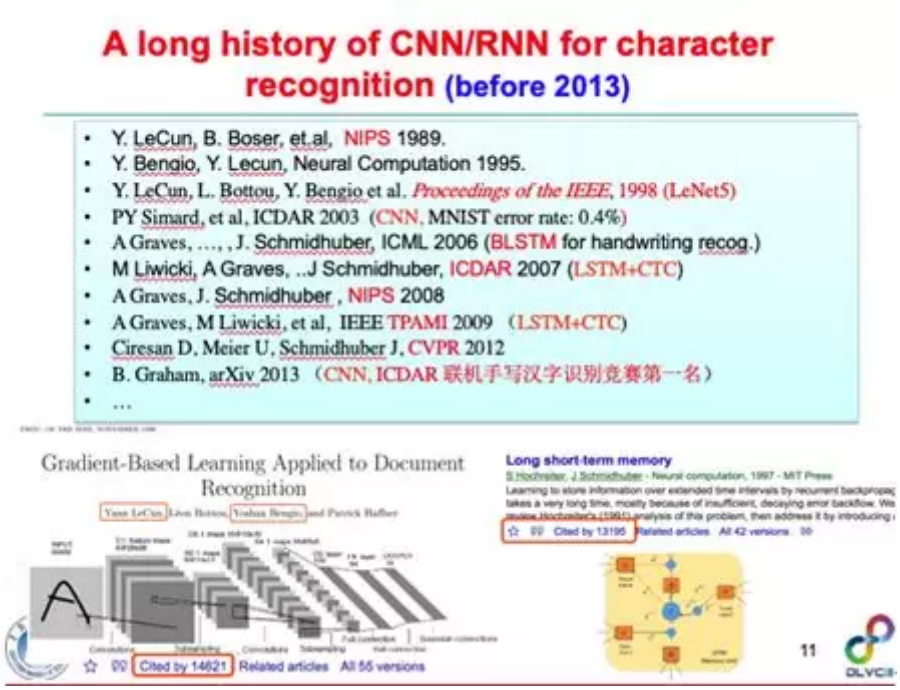
在讲场景文字识别之前，我猜在座有很多老师和同学在这个领域做了很多年的、很有经验了，可能也有一部分同学老师不是做这个方向的，所以我先把一些基本的概念及此领域发展历史简要给大家介绍一下。

传统来讲文字识别是一个典型的模式识别问题，跟其他模式识别一样，传统文字识别主要包括预处理、特征提取、分类器等几个步骤，在2012年以前主流的框架是这样的，典型的代表性的工作我列举在屏幕的下方。这个框架很繁琐，每一个模块都需要很好的设计才能达到很好的性能。在今天深度学习时代虽然这个框架基本上被大多数人抛弃掉了，但是近年来我们看到仍有一些学者还是遵循这个框架做一些工作，比如典型的例子是两年前（2017年）谷歌的研究人员发表在TPAMI上面的文章。

自从2012年深度学习兴起了以后文字识别的技术就变得比较简单了，之前做文字识别我们招一个学生培养他入门他要花一年多才可以把特征提取、分类性设计等相关的基础技术掌握，但是今天没有这个问题了，我现在招研究生面试的时候，有时会出一个文字识别的小题目给本科三、四年级的同学做，大部分同学1到2个月的时间就可以重现一个不错的文字识别的算法。因为深度学习技术的普及，文字识别特别是单字符识别的解决变得非常简单。当然近年来也出现了不少文字检测与识别的新方法，代表性的工作我这里也列了一些。

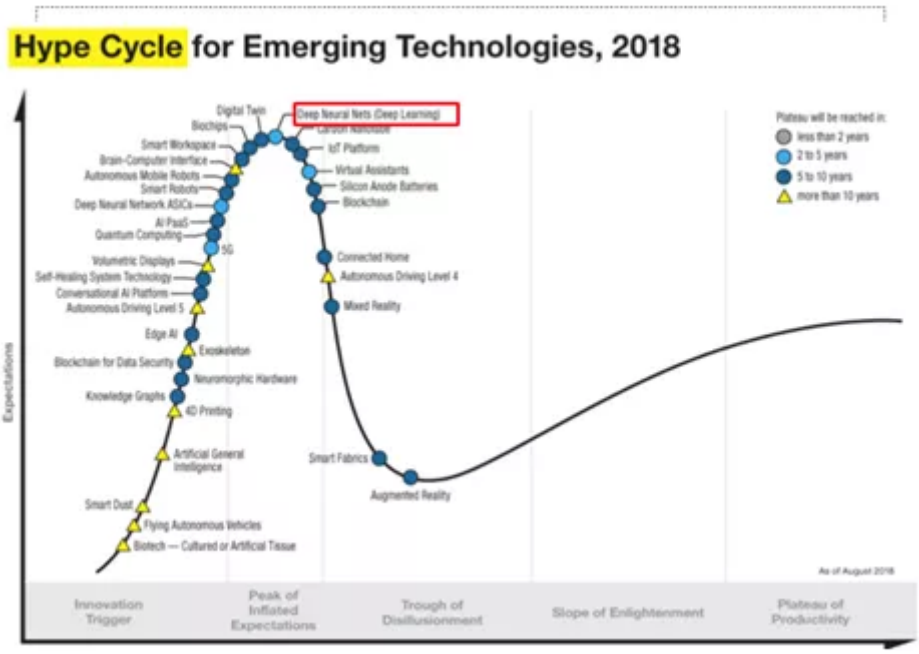
这里值得一提的是，深度学习在文字识别领域的应用有很长的历史，在深度学习还没有火之前，在文字识别领域当中一直在采用像卷积神经网络、循环神经网络的技术来解决文字识别领域中的一些问题，比如构建了美国邮政编码的手写识别系统。在2003年微软的剑桥研究院的学者第一次用CNN再加上数据增广的思想把识别率做到了99.6%，这个精度此后近十年时间没有几个方法能超越（当然现在已经可能很轻松做到99.7%以上了）。LSTM、BLSTM在文字识别领域很早就用来做手写英文的识

别和手写公式的识别，只不过那个时候没有GPU，所以做实验非常困难，我们的学生在2009年做手写英文识别时，做一次十几万数据规模的实验大概要三个星期左右才能出结果，所以当时没有同学愿意搞循环神经网络这个方向，直到后来有了GPU及各种不同的开源工具。其它一些比较典型的工作，比如LSTM与CTC结合是序列识别中比较经典的算法，这个在很早之前就有人提出来了。这里列出的是两篇比较经典的文章，引用率都很高。



后面的发展大家可能都知道了，深度学习进入到学术界被逐步认可变成主流是从2012年开始，进到公众的视野可能是从Google的AlphaGo开始的，全世界许多媒体报道了这次人机围棋大战，大家才意识到以深度学习为代表的人工智能原来可以有这么强大。后来在2018年的时候有人提出了一个观点叫深度学习技术将来是无处不在，当然学术争鸣是不可避免的，其中一个批评的声音是说深度学习就是调参工程、是炼金术，还编了一些很好听的顺口溜，比如“性能不够 加层来凑”、“数据不够模型补，模型不精数据上”等。在学术界总是有不同的声音，我觉得也挺好，我后面也会讲一下深度学习确实是存在不少不足及问题。但这些问题我们应该理性去思考。关于深度学习的讨论还曾经争议到了《Nature》上，在2016年《Nature》上面有一篇评论文章专门讲深度学习的黑盒子问题，但是它最后有一个观点我很赞成，它说虽然某些领域因为深度学习的不可解释性所以带来不确定问题，但是我们不应该过分地苛责这个问题，就像人的大脑一样，虽然我们对大脑的机制还不完全清楚，但并不妨碍我们每天都在使用大脑。

这是去年Gartner发的技术成熟度的曲线，大家可以看到深度学习还是目前最热门的技术之一。



我后面要介绍的场景文字检测也是从深度学习的视角来做的一些介绍，再结合我们的实践经验给大家介绍两种新方法。

首先，我先介绍一下场景文字检测与识别的一些近况及发展趋势。场景文字检测识别目前来讲大概分为场景文字检测、场景文字识别、以及端到端文字检测和识别三个主要的方向。其中文字检测方法主要是包括基于文本框回归的分类、基于分割的回归、以及分割和回归结合的方法。过去两三年它的发展趋势，从早期2015年以前是以水平的矩形框检测为主，后来发展到多方向的矩形框，再到2017年开始有人做任意的四边形，再到这两年任意曲线文本的检测，大概是这样发展的趋势。矩形框的文字检测基本上解决得很好了，只要你有一定的数据量基本上可以做到比较不错的识别性能。前两年任意形状（例如曲线）文本检测还是一个比较挑战的问题，但是这两年有很大的进步，比如像今年CVPR、IJCAI发表的与文字检测相关的文章中，80%基本上都涉及到任意形状的文字检测。

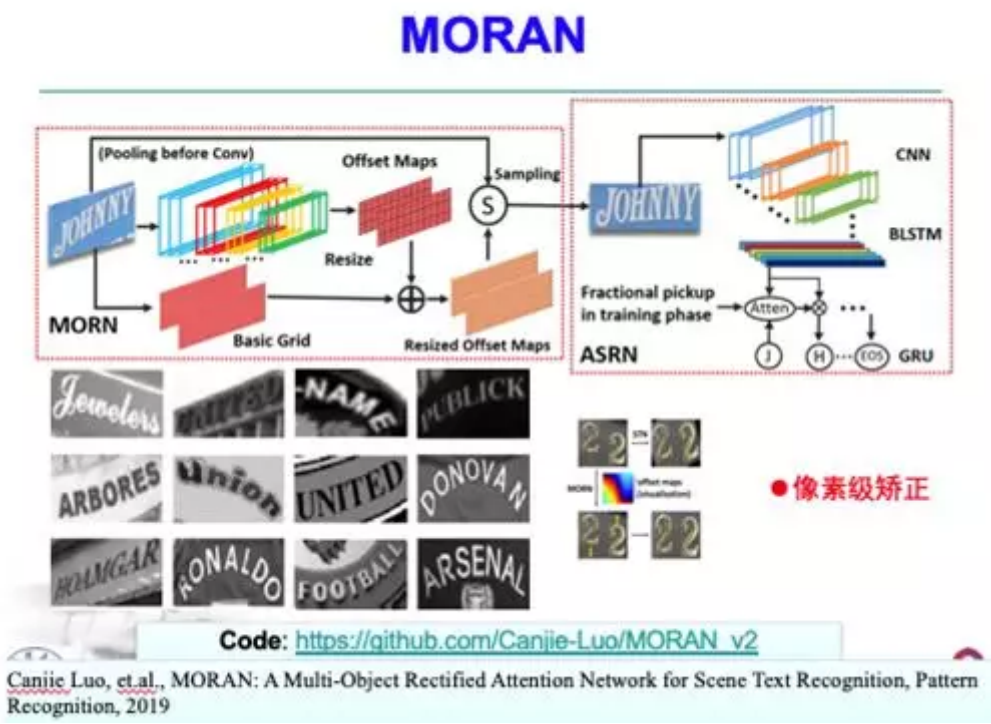
场景文字识别的传统方法很复杂，但在2015年之后基本上主流的方法是基于两套思路，一个是基于CTC的方法，尤其是CTC和神经网络的结合，典型的代表方法是CRNN;另外一套思路是基于Attention的方法，典型的工作是从2016年开始，再到今年华中科大的廖博士发表在IEEE TPAMI上的论文，都是基于Attention的机制。

场景文字检测的挑战在哪里？首先它的方向是多变的，第二个有很多不规则的文本，比如在街边看到的广告有很多弯弯曲曲的文本，其次一个图片里面文字的尺度是有多样的，比如说有些字很大，有些字很小，同时一个算法要同时解决这些问题很难，此外还有标注歧异性的问题、检测的完整性问题等等。有些问题得到了不错的解决，比如多方向、任意方向的文字检测是可以的，有些问题并没有得到特别完整的解决，比如说尺度问题和不规则文本的问题还没有得到很好的解决。

场景文字识别主要的挑战有哪些呢？比如变形文字、不规则文本、曲线文本、字体的多样性、形状的多样性、自然场景中拍的照片有各种背景的干扰、对焦、图像模糊、遮挡等等问题，目前还有不少问题还没有解决得很好。如果大家去做这个领域的研究，你去分析一下这个领域当中面临的挑战及困难在哪里，实际上可以启发我们做研究的思路，如果能想一些方法去解决好目前面临的问题和挑战，就能够带来一些不错的研究成果。

举例来说，比如不规则文本的识别，这是过去两年学术界比较关注的问题之一，就是指文字形状不规则（例如弯弯曲曲的广告文字）如何来进行有效识别？很多学者想了不同的招数，开始最直接的想法就是对文字进行矫正，比如早期CVPR 2016用STN的方法把它矫正过来，后面发展到TPS的变换找一些控制点，对控制点进行回归学习后把文字进行矫正，包括在像素级别做矫正等等，都可以在一定程度上解决此问题，这个是矫正的思路。还有一些学者是从两维attention的角度解决这个问题，也提了很多方法。还有一些学者是从字符级的角度解决，比如说这个文本行不论怎么变形，字符级的旋转有限而且相对容易进行识别，如果可以在字符级设计一个很好的检测器及分类器，我们就可以很好的把不规则文本识别这个问题解决。

我刚才列了很多这个领域的挑战和问题，我认为还有大量的问题没有解决好，大家把这些问题好好想一想，还有很多做研究的空间。



下面我简单回顾一下场景文字检测与识别此领域的一些代表性方法。在场景文字识别当中最有代表性的一个方法是就是华中科技大学白老师团队最早在2015年做的CRNN模型（后正式发表在IEEE TPAMI2017上），记得2015年我请他去我们学校做报告，当时白老师做完报告后不到一两个星期就把这个代码开源了，后来我学生把它改写成了C++的代码，发现确实很好用。今天不少公司做的OCR引擎都采用了这个框架。

这是解决不规则文本识别的一个典型，比如说文字是弯曲的，你要把这一串文本行识别出来，一个最直接的思路就是学一个变换把文字矫正过来，或者找一些控制点回归以后把它给矫正过来。代表性的工作包括RARE（CVPR 2016论文）以及ASTER （IEEE TPAMI2018论文）。这是我们实验室做的一个工作，是从另外一个思路去解决这个问题的，与学变换函数不同，我们直接从像素点上去学不规则文本它的变形，然后从像素级别上进行纠正，这样通用性更强。识别效果当初可能是最好的，代码我们也开源了。

在今年7月份，我在arXiv上还看到一篇2D-CTC的论文，以前CTC只能从一维找最佳的路径，该方法增加在高度方向的概率分布图以及一个路径转移图来解决二维CTC问题，想法还是很有新意的（虽然还不能说是真正的二维CTC）。跟Attention比它最大的好处就是速度快，比Attention快很多倍，性能还更好。

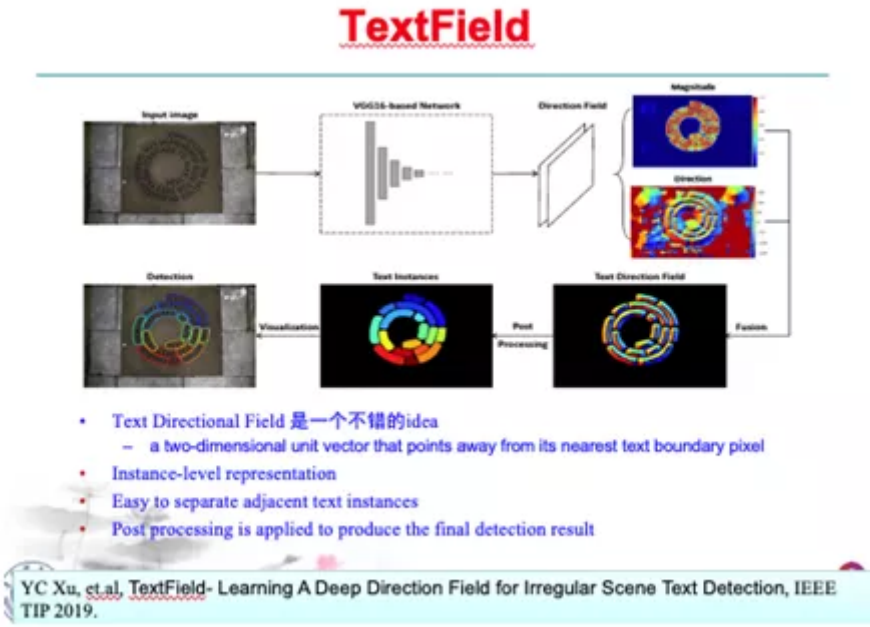
刚才我举了场景文字识别中几个代表性的工作，我下面简要讲一下文字检测中一些代表性的工作，这个TextBoxes模型早期很出名（因为它是早期开源的检测代码之一）。这个工作其实就是SSD的模型基础上改了一下Anchor，然后针对文字做了处理就变成了Textboxes的模型。后来此方法进一步改

进为可以做任意多方向的四边形文字的检测，同时可结合CRNN网络做成端到端协调训练的解决方案，这个TextBoxes++发表在IEEE TIP2018上。

这个是我们实验室在2017年做的一个工作DMPNet，我们当初关注到任意四边形的检测问题，提出了一个实现任意四边形文字检测的方法，当初以这个模型为基准进行适当扩展及改进获得了ICDAR2017年MLT多语言文字检测的冠军。

同一年也有不少人关注到任意多边形的文字检测问题，最典型的代表就是EAST模型，这个模型它最大的特点就是简单、有效，速度也比较快。它是基于FCN的框架去做文字检测，可回归多方向的矩形框、或者任意四边形，这个模型后来被OpenCV3.0采用变成了一个标注库。

这是华中科技大学许老师等做的工作，提出一个TextField的概念，就是文字方向场的概念，传统基于分割的文字检测方法有一个很大的局限性就是对密集文本无法有效区分开，他们提出一个文字方向场，基于像素做回归，然后通过后处理组合成一个文字条，对于弯曲特别离谱的文字都可以检测出来。



去年到今年有很多学者关注不规则文字检测这个问题，这是今年CVPR的一篇论文，我觉得它的原理很清晰明了，而且方法也很简单有效，效果也非常好的。另外，CVPR2019今年百度还有一个工作，这个工作取得了一个很好听的名字叫“多看你一眼”（ Look More Than Once， LOMO），这个工作实际上是想解决两个问题，一个问题是在直接回归方法的基础上往前走了一步，解决长文的检测性能不够好的问题，提出了一个IRM模块去解决长文检测；另外一个问题是弯曲文本的检测，提出了一个SEM模块，回归一系列几何属性，比如说字符的中心线，每个像素到字符边缘的距离等等，通过一些简单的后处理可以把任意形状的文字检测出来。此方法几个模块端到端可以训练，不需要很复杂的后处理，简单有效，也是一个不错的工作。

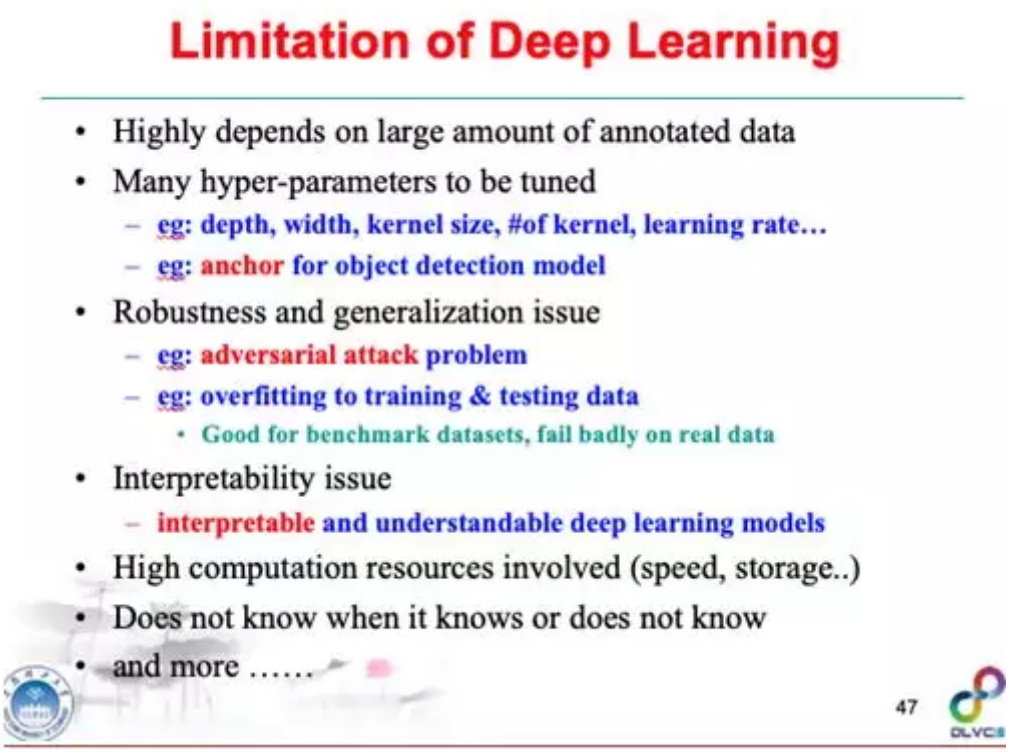
今年还有一个比较新的工作，是8月份出来的一篇文章，模型名称叫做PAN，就是像素聚合网络（Pixel Aggregation Network），这个网络与去年提出一个叫PSENet模型的团队大致是同一批人，PAN这个工作最大的优点是快、同时效果也非常好。

端到端做的学者相对比较少，但是过去两年也有不少工作报道，可能是未来发展趋势之一。最早在ICCV 2017年就有学者开始尝试解决此问题，当然端到端也不太容易做，之前的一些端到端的工作还是伪端到端的方法。CVPR 2017商汤提出了一个FOTS的模型，当初取得了非常不错的端到端识别效果。这个工作是去年ECCV2018提出的Mask TextSpotter模型，虽然是基于Mask RCNN进行改进，但它提出把原本的Text/Non Text二分类改进为多分类，再加上适当后处理变成了一个简单、有效的端到端方案，我觉得还是一个挺大胆有效的想法，我们去年有同学也把它复现了，效果还是不错的，今年此论文作者又加了一些新东西，包括Spatial Attentional Module等，使得这个模型相比之前的版本不需要做字符级的标注也可以做端到端的训练，而且效果挺不错，最新的论文发表在IEEE TPAMI上。

这个表总结了一下场景文本检测领域当中现在的10来个数据集目前最好的结果，大家可以看到一些常规的数据集，包括多方向矩形框或者四边形的文字检测差不多解决得很不错了，但是有一些数据集还是很难的，比如中英文混合的，垂直水平曲线混合的，包括我们做的SCUT-CTW1500弯曲文本数据集，它的性能指标还有待进一步提升。这张表格是总结了规则文本识别的现状，一些常规的英文字符识别率指标可以做到95%以上了，这几个数据集进一步刷已经没有什么太大意义了。不规则文本还值得去做一下，比如像Cute80、SVTP、IC15等数据集上的性能指标离产品要求还有很大的距离，一个好的产品最好有90%、95%以上的识别率，最起码要85%以上才可以用。

另外，我们实验室最近总结了三个资源，就是场景文字检测的资源（https://github.com/HCIILAB/Scene-Text-Detection）、场景文字识别资源（https://github.com/HCIILAB/Scene-Text-Recognition）和端到端文字识别的资源（https://github.com/HCIILAB/Scene-Text-End2end），这个资源我们昨天把它开放出来了，也写了三篇公众号的文章，大家感兴趣可以关注一下我们CSIG文档图像分析与识别专委会的这个公众号，我们会不定期给大家推荐一些论文、代码、数据集等资源，欢迎大家扫码关注。

如果有一些同学是刚刚接触这个领域的，有一些比较好的综述性的文章给大家推荐，这一篇TPAMI 2015年发表的一个传统方法综述的文章。这一篇文章主要是做视频文字识别检测的综述文章。这是《自动化学报》去年的综述文章。另外，去年还有一个最新的综述，是关于深度学习时代场景文字检测和识别的现状，大家感兴趣可以去找来看一下。



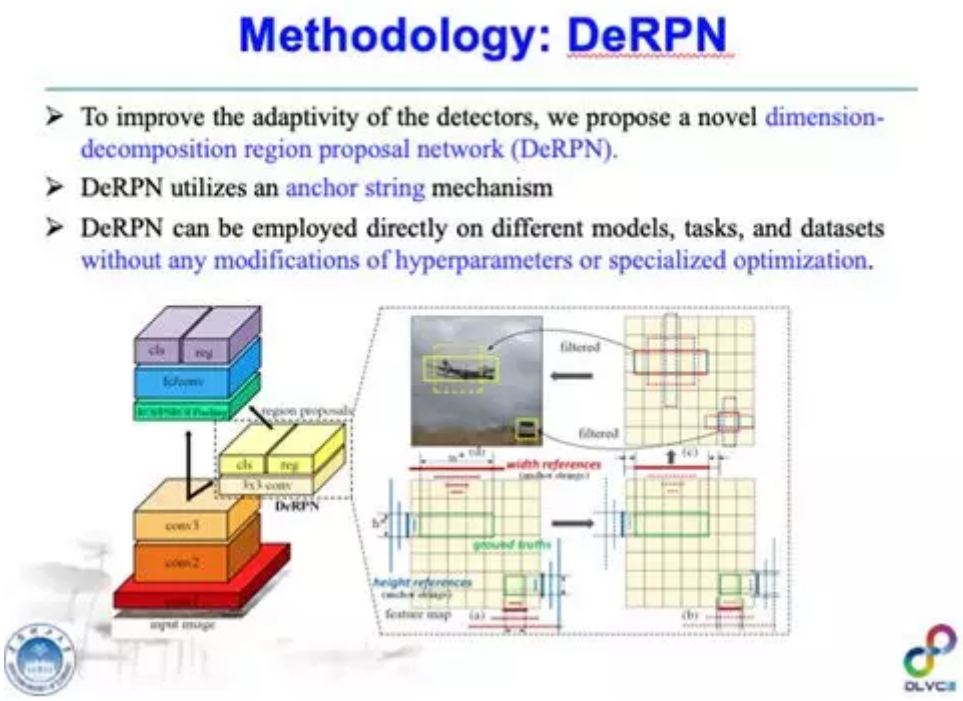
虽然深度学习是场景文字检测和识别主流的方法，但是它也存在一定的局限性，首先它极大的依赖于足够好、足够多或者足够高质量的标注数据，这是深度学习的优点也是缺点，优点就是说我们解决一个问题，例如图像识别问题，只要给足够多的高质量数据，数据的分布也足够有代表性，那总能训练出一个足够好的识别引擎，这个没有什么难度。虽然它是优点但是也是缺点，没有数据的情况下你的模型就傻眼了。第二个问题我还想说一个观点，我们今天不少深度学习的模型它都可能是在训练集和测试集上过拟合了，虽然测试集和训练集的数据是分开的，但是一个模型针对同一个问题在这样的数据集上也许可以达到很好的效果，比如我的OCR在中国银行的票据识别做得效果很好，换到中国农业银行可能版式、字体、背景等改了一下这个模型可能就没那么好了。一个模型在某个开源的数据集上做到了很好的效果，用这个方法直接换到另外的数据集上也许就不是最好的，甚至是挺差的。因为不少模型是针对某数据集来调参数优化的，它极大依赖于数据，深度学习它有没有学到本质的东西，这个问题还值得探讨。因此有时候深度学习的鲁棒性和推广性还不够好，而且它容易受到对抗样本的攻击，你在样本中加一点噪声它可能完全被识别成了其他的物体了。

还有深度学习大家比较批评的是说它是一个调参工程，是炼金术，有大量的参数需要去调，早期的时候确实是这样的，今天因为有很多成熟的技术、成熟的模型，其实也不需要那么多的调参工作。对于检测问题来讲，基于回归的检测框架不少需要用到anchor，它的参数设计也很重要，如果anchor的参数设计不好，性能可能会很差。

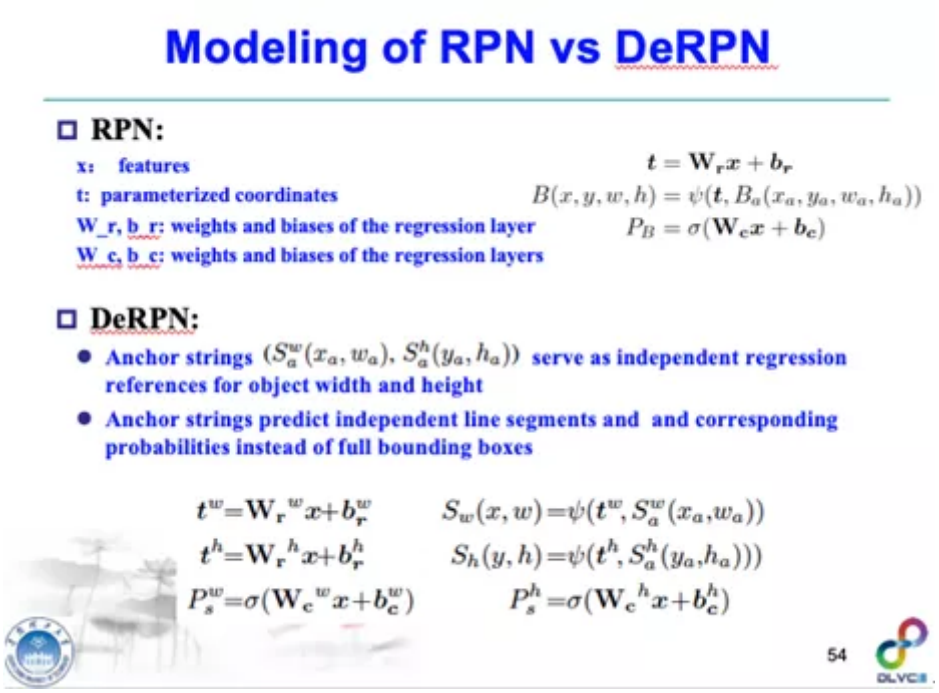
另外，深度学习的可解释性比较差、计算复杂度比较高，它不知道自己知道，它也不知道自己不知道，缺乏一些常识，这些都是目前深度学习技术存在的不足及问题，都值得去解决。

虽然深度学习有这样一些局限性，但是它确实可以帮助我们解决很多问题，所以我们可以看到过去两年在场景文字检测和识别中基于深度学习的方法是此领域绝对的主流。针对深度学习的某些局限性问题，我们也做了一些工作，第一个工作就是解决anchor调参的问题，论文发表在AAAI 2019上，这是一种解决anchor调参的新方法，我们把它命名为DeRPN。

这个工作的背景是什么呢？通用的计算机视觉当中物体检测的方法有一类模型是anchor based的方法。基于anchor的方法有一个问题是需要对anchors很好的去设计，当然可以去穷举、可以设计很多很多不同的anchors，但是anchors越多，计算复杂度越高速度越慢。而且针对不同的问题，比如计算机视觉通用的物体检测可能是一种anchors，你要做文字检测可能又是一种anchors，你每次都要做一些参数工程化的工作，挺麻烦的。所以我们就想能不能把anchor的设计分解一下，以前是anchor box，是一个2D维度，如果把两维的问题分解成一维问题，这个问题要简单得多，参数空间也少很多。这种把两维anchor box分解为一维之后，我们取名叫anchor string。它的原理我等一下讲，它有一个很大的好处是，这种anchor设计方法基本上可以用在所有基于anchor的检测框架上，而且此方法在anchor设置上不需要调参数，也不需要做特殊的优化，在不同模型不同任务上基本上都可以达到很好的效果，我们做了不同的实验，我们针对场景文本、计算机视觉中的通用物体的检测，我们都没有改过任何参数，同一个模型直接在几个不同的数据集上都可以取得不错的性能。



这是我们提出的基于anchor string的方法DeRPN，我们把传统anchor box分解成在X的方向及Y的方向两条独立的边（anchor string），这是一个维度降维的思路，把两维的box变成一维的string，每一个string用它的长度及坐标去建模，这样anchor的参数设计变得很简单，比如说你可以把它变成一个等比的数列（eg 16、32、64、128...），我们在论文中提到，如果设计一个等比数列的情况下回归精度的误差是有界的。而且此方法可以很好的去解决多尺度的问题（这在物体检测当中尤其是在场景文字检测当中是一个很重要的问题）。

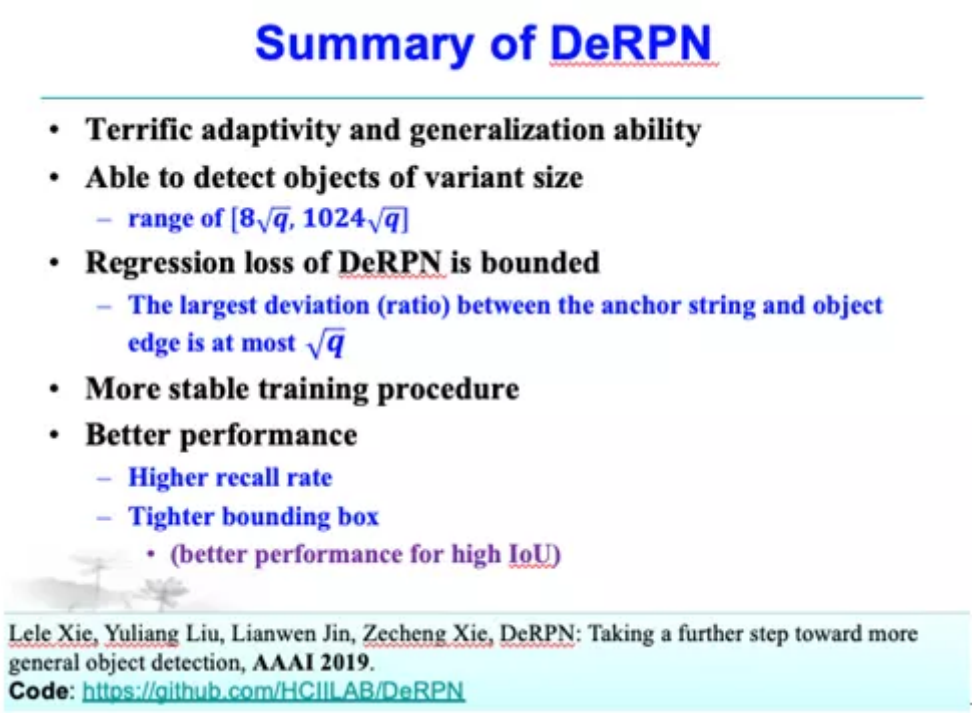


当然这个想法看起来比较简单直接，但具体实现时还是要解决一些问题，例如训练样本标注的生成问题，即解决anchor string的Label生成的问题，第二个问题就是对回归得到的不同一维strings，如何两两组合还原为物体或者文字的两维位置（Bounding Box），针对这些问题我们提了一些简单有效的方法。

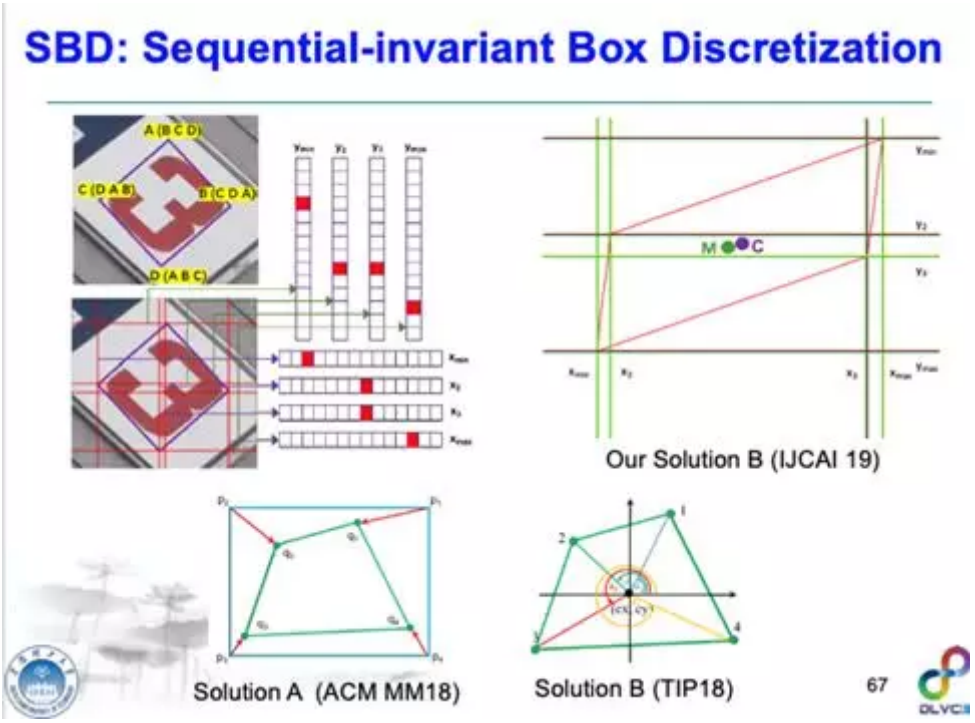
我们做了大量实验验证，包括场景文字的检测、通用物体检测等，在这些实验当中我们不做任何定制化的参数调参，我们用相同的网络、相同的参数解决不同数据不同检测任务。在MS COCO及VOC等通用物体检测数据集上，跟传统的RPN的检测网络比，在高IoU的情况底下我们的方法比RPN明显好，甚至好了十几个点，这说明我们这个方法它检测出来的框是更加完整，传统的方法IoU只要达到0.5就认为是对了，但也许没有框完整，我们的检测框更完整，当然整体的指标我们也是好的。
在场景文字检测当中，在ICDAR 2013数据集很轻松可以做到90%的检测准确率。在MS COCO Text数据集上，与之前不同的anchor设计方法（例如有手工调参的、Kmeans聚类的、coco-type等），我们的方法明显要好很多的，关键是它不用怎么调参，一个网络走天下。

这是一些可视化的结果，这里有四行，第一行和第三行是我们的结果，第二行和第四行是传统方法的结果，传统的方法有很多框都没有检测出来，我们把它们检测完整了。传统的方法检测框有时框多了背景、有时漏框了很多文字，我们框得更加紧凑完整一点。

总体来讲这个方法它有很好的通用性和自适应性，它可以适配不同的物体检测的尺度，我们证明物体的尺度变化只要在这个范围之内，我们就能够把它检测出来。另外我们也大致算了一下，它的回归Loss是有界的，最多就是根号q，这篇论文中q通常取2，所以回归的误差比较小，此外此方法训练的过程更加稳定，性能更好，框得会更紧更完整。这个工作代码我们今年也把它开源出来了。



受到此工作的启发，针对多方向的文字检测，以及标注点顺序歧义性问题，我们提出了一种把检测框分解为四个关键点及相应的关键边（Key Edge）的思路，但不是在string这个框架下去分解，我们是在Mask R-CNN的基础上增加了一个SBD的模块，在任意方向的矩形框去找到一些关键的边。这个工作原本的初衷主要是要解决标注歧义性的问题，比如说这里是一个文字框，我们常规是ABCD这样标的，如果标成了CDAB好像也可以，但第一个点顺序标得不一致的话可能会对检测效果产生很大影响，我们做过实验，随机增广或打乱的标注点对一些检测器它的检测性能可能会降到十个百分点。



所以我们想这些标注的点有没有一些点是不变的？确实有一些点是不变的，比如说检测框它的X方向和Y方向的最大值和最小值的，这四个点无论你的标注顺序怎么变，它都是不变的。在学术界里面这个问题被大家关注也很久了，例如ACM MM18的论文以及TIP2018的论文都提到过此问题，但是这个问题仍然没有解决好（比如说第一个点在横坐标上面，就是X坐标附近，你稍微变一个像素它有可能从第一个点就变成第四个点了），细节我就不讲了，所以标注的歧义性问题并没有解决得很彻底。另外，在我们提出的这个基于关键点及关键边的检测框架中，我们还把回归组合的问题变成一个分类的问题，所以检测框的置信度比传统方法的置信度更加可靠，细节我没有时间讲了，大家可以看我的文章。此方法它的推广能力也很强，比如我们用相同的网络去做多方向卫星船舶图像的检测，不用调参，我们没有改任何的参数，只训练了两个多小时就可以比之前的结果好一大截，所以此模型的推广能力还是不错的。我们今天参加了ICDAR2019 ReCTS场景文字检测比赛，基本上是基于这个单模型，把主干网络稍微改进了一下，拿到了检测任务的冠军。

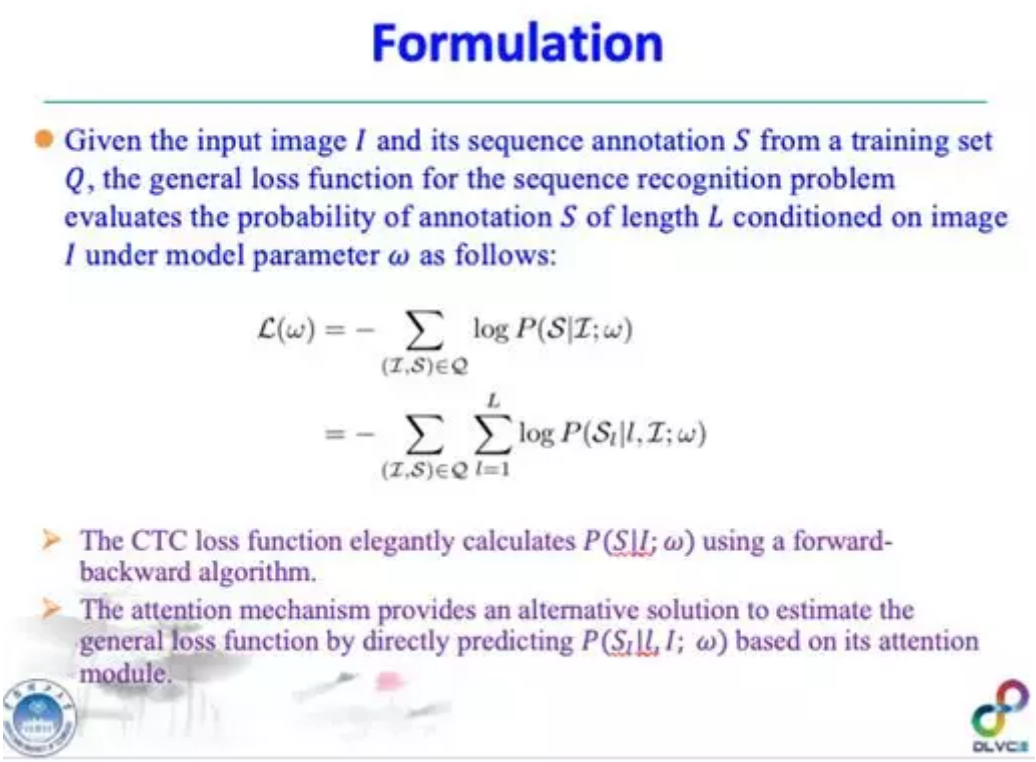
下面介绍一个文字识别的工作，是文本行识别的一个新思路，我们把它叫ACE（Aggregation Cross-Entropy），是从一个全新的角度去解决文本行识别问题。

先介绍一下研究背景。序列识别经典的方法是CTC，另外过去三年attention机制变成了主流，当然如果大家做工业应用的话，可能会发现attention机制虽然很火，做英文识别效果确实也非常好，但是做中文尤其是中文长文本识别可能是干不过CTC的。这两个机制都有一定的局限性，比如CTC最早是在语音识别领域提出的，提出十多年了，这十多年来也有很多人注意到了它的一些问题，比如说前向后向算法比较复杂，计算量和存储量是跟序列长度及类别数相关的，长文本大类别情况下计算复杂度比较高，另外CTC 无法解决两维的弯曲文字的识别。Attention机制是从自然语言处理领域来的，2016年开始用到文字识别的领域，attention有一个很严重的问题就是对齐的问题，就是attention漂移的问题，如果一个长文本中有一个地方没有注意到，或者某个噪声干扰到它了，attention错了，后面会错得很离谱。这个问题很多学者意识到了，但可能还没有很彻底的解决。另外，attention它的存储量和计算复杂度是很大的，特别是类别数很大的情况下参数增加是很多的。

序列的解码有没有一些新的思路去做呢？我们提的方法很简单，传统的方法在序列解码的时候不是要预测每一个序列当中每一个字符的概率吗？我们把预测字符的概率转化为预测字符出现的次数的概率，因为直接去估计这个字符在某一个位置的概率是不太好准确计算的，但如果能大致的估计它出现的次数，相对就简单一点，这个方法的核心思想就是把预测字符分类的概率转化为预测它出现次数的概率，这样做它还会带来一个很大的好处，就是训练过程跟标注的顺序完全无关，你只要告诉我出现了哪些字，每个字出现了多少次，训练样本怎么标都没有关系。

这个想法从数学上来看基本上还是合理的，我等下给大家解释一下。这是一个传统的序列解码的数学建模的模型，给定一个输入的图像和模型参数，我们的目标把这个序列预测出来，通常会定义出一个Loss，然后转化到每一个时间点字符出现的概率的loss，CTC和attention都可以估计这个概率，但直接估计这个概率是很困难的，我能不能转化成估计这个类别出现次数的概率。为什么这里可以大约

进行转换？因为基于Softmax的预测概率值往往对于真实类别是倾向于1的，因此预测概率与出现次数的概率这两者差不多是相等的。



所以这样一来这个问题就变得简单很多，设计好Loss Function然后去优化网络参数就OK了，最直接的方法是用基于回归的loss。但是对网络模型求导的时候会出现梯度消失的问题。因此我们用交叉熵来度量他们分布之间的距离，这样你在求导的时候它的梯度弥散的问题就不存在了。这个方法还可以平滑推广到两维甚至多维的情况，因为我们计算类别出现次数的概率，因此两维的时候你两维加起来并稍微用长和宽归一化一下就好了，推广到两维完全没有压力。要实现这个方法也很简单，就是这四条公式，几行代码差不多就可以实现了。

我们做了一些实验来验证ACE方法的有效性，包括场景文字、手写体文字识别，另外还有一个计算机视觉当中的数物体的数量的实验。

先看一下我们两种解码的方式的对比，即回归和交叉熵的方法，我们发现交叉的方法好很多。跟TPAMI 2018的工作比，性能相当。但是如果在样本标注顺序随机打乱的情况下，我们的方法就比较有优势，如果随机打乱样本的标注顺序，传统attention或者CTC的方法性能就急剧下降，我们的方法得到的性能是不受任何影响的。另外在场景文字识别当中跟之前的方法比，特别是不规则的文本识别我们做到不错的结果。另外一个实验室手写体的汉字识别，这其实是一个挺难的问题，尤其是汉字类别数很多，与之前主流的方法相比，我们的方法在没有语言模型的情况下明显比较好，有语言模型的时候是相当的。第三个实验做了一个Counting的问题，就是给你一张图数里面有多少个物体，用我们的方法很轻松可以做到State-of-the-art。

另外，我们也实际测了一下各种方法的计算的复杂度和存储复杂度。CTC在小类别的情况下还是挺好的，它又快存储量也不大，attention相对来说就很慢了，attention比CTC要慢差不多20倍，但我们的方法比CTC快差不多30倍，所以我们是最快的。如果针对大类别的序列识别的问题，比如说手写汉字识别的问题，有7000多个类别，CTC要47.8M的存储量和十几毫秒的计算时间，而attention要140多M、85毫秒的解码时间，相比而言我们的存储量少了10倍到30倍，速度快了160倍到800倍，所以这方法在速度和存储量上是有很大的优势的。

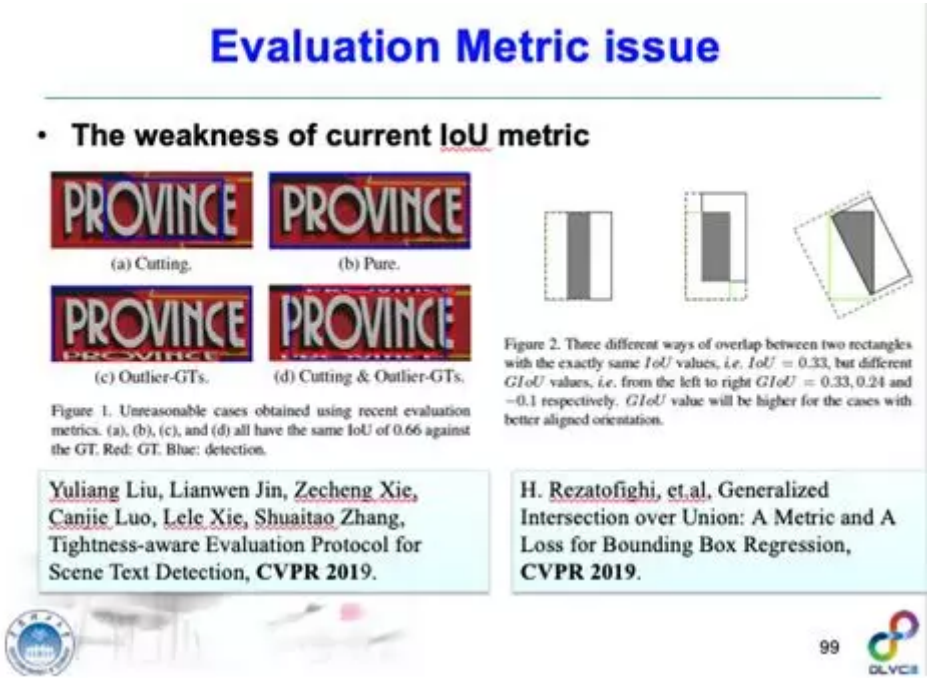
Complexity Comparison

Table 5. Investigation over parameter (Para), runtime memory (Mem), and speed (Speed) (in units of MB, MB, and ms, respectively) of CTC, attention, and ACE.

Method	37 classes			7357 classes		
	Para	Mem	Time	Para	Mem	Time
CTC	no	0.1	3.1	no	47.8	16.2
Attention	2.8	6.6	78.9	17.2	143.6	85.5
ACE	**no**	**0.02**	**<0.1**	**no**	**4.2**	**<0.1**

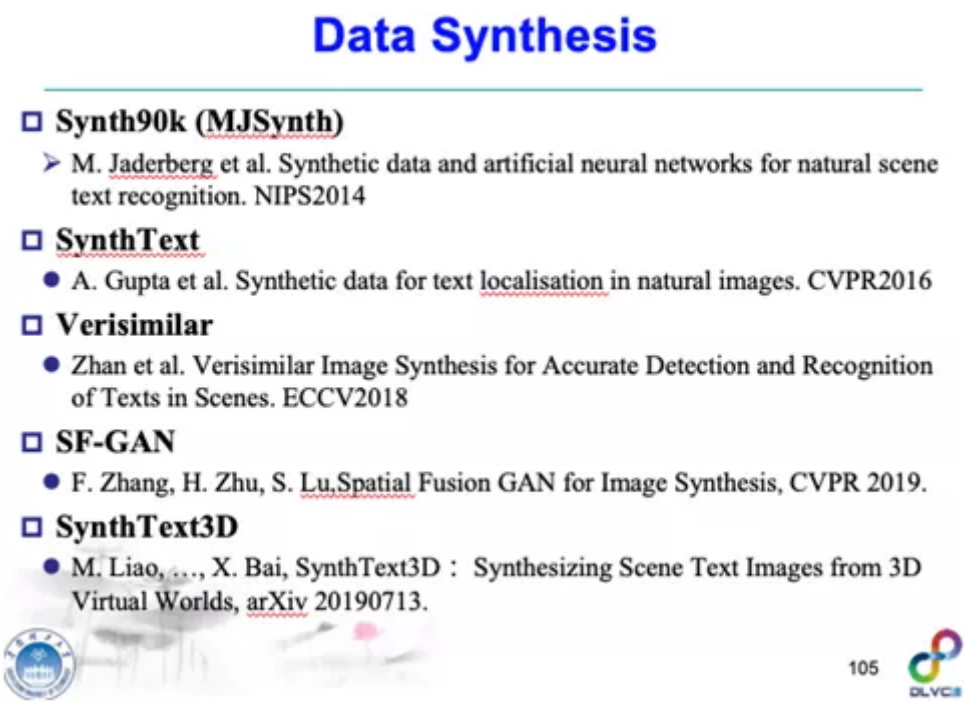
总结一下这个方法实现起来很简单、速度快、存储量小、用起来也很方便（只把CTC换成ACE就OK了），而且它还有一个副产品，就是它是跟标注顺序无关。我们这个工作也开源了，大家感兴趣可以去试一下。

下面我再花几分钟讲一下场景文字检测识别领域从去年到今年一些发展趋势及存储的一些问题。其中一个趋势就是说场景文字检测的评测标准是一个有待探讨的问题，现在场景文字检测大部分还是利用计算机视觉中的检测标准来评估，比如说用IoU大于等于0.5被认为是正样本，这个标准有时无法评估场景文字检测器的优劣，比如左边的图中四个检测框它的IOU都是0.66，但是有一些显然没有检测完整，有的会额外包括很多噪声，这些对识别影响都很大，但目前的评测标准看不出差异来。



所以IOU的标准怎么解决是一个重要的问题，我们提出了一个TIoU的标准，考虑到了检测得不完整、或者不紧凑情况下的惩罚项，当然这个方法目前还是比较经验化的，我感觉这个问题也还没有彻底解决好，如何更客观公正的评价检测框的紧凑性、一致性和完整性，还值得进一步想一下。TIoU这个代码已经开源了，大家感兴趣还应该往前发展一下。当然文字识别的评测标准也有一些争议，比如这篇文章（好像是ICCV 2019录用的论文），我认为其贡献看标题及摘要差不多就够了。

还有一些新的趋势是做端到端可训练的场景文字的擦除，这是今年AAAI 2019的一个方法，数据已经开源。另外，数据的问题也很重要，在没有人力物力标注足够多数据情况下，一种解决办法是做数据合成，这里有不少代表性的方法，这张PPT给出了一些合成结果，看起来还是挺像的，而且也有用。但是真实的数据还是很重要的，这篇AAAI 2019文章做出来的效果非常高，特别是在CUTE80直接秒杀其他所有的方法，后来我们仔细看了一下这篇文章，发现原来他用了真实的数据，如果不用真实的数据只能做到83%左右，所以真实的数据还是很重要的。



解决数据不足还有一个思路是做数据增广，数据增广怎么做得更好更合理，比如说谷歌他们今年CVPR 2019有一个方法AutoAugment的方法，就是用强化学习的来找最佳的数据增广策略。

其他的发展趋势我个人认为还有不少问题值得解决，比如说端到端的问题没有彻底解决好，标注的歧义性问题还没有解决好，多尺度问题大家意识到了，今年有不少文章尝试着要去解决这个问题，但可能都还没有彻底解决好，另外场景文字检测文字级别及文本行级别的几何属性学习可能也是一个趋势，这两年我们看到了很多论文从不同角度关注到了此问题。

上面列举了一些发展趋势及问题，可能是值得大家关注的一些研究的点。
我的报告到这里，谢谢大家！

嘉宾：大家下面可以提问。

提问：金老师您好，我想请教一个问题，非常感谢你的精彩报告，也对近几年的工作进行了一些梳理。我想问一下，你对文字的识别和主动学习或者在线学习方面能不能有一些结合或者前沿性的介绍？

金连文：对于主动学习和在线学习领域我不是很熟悉，但是我觉得是可以考虑去做的，你的产品线上肯定要考虑增量样本学习的问题，或者你可以把人的因素设计到学习框架里面去（Human in the loop）。但是怎么把它设计得很好，我还没有看到好的工作汇报。这方面我们没有经验，但你在解决实际的应用系统当中肯定是要去思考解决这个问题，新的数据新的场景来了不能又从零开始重新建模，但是用什么样的框架去做，这个我还不太清楚。