AIoT

这一次唐剑老师为我们带来的是关于AI+IoT的报告，将当下两大热门的领域相结合。其实细心想想AI与IoT似乎本来就具有极好的结合运用可能。在5G技术是的物联网的落地速度加速的当下与未来，似乎更具挑战的课题是如何运用好物联网随时随地都在产生的数据。我们知道在人工智能的训练之中很重要的一环是通过已有的数据去检验算法的精确性并预估投入实际问题解决的可行性，如此需要拥有大量的数据作为支撑。而现在物联网则提供了获取海量某一特定领域的数据信息的契机，甚至在未来有需要时为多领域人工智能课题提供范围更广的数据集。反过来，人工智能的发展也将使我们对于物联网所产生的海量数据拥有更为智能的应用。正如唐老师所言：“如果说5G是使IoT落地的关键技术，那么引入AI就是能最大程度挖掘IoT的潜能并使其真正变得智能，变得无所不知，无所不在的技术。”

目前由于并没有真正开始学习人工智能有关的课程，每当提起AI的时候总是绕不开深度学习，计算机视觉，自然语言处理这三个方向。而这次唐剑老师为我们拓宽了视野，向我们介绍了日前正在逐步兴起的另一个研究方向：深度模型压缩。当下我们对物联网所产生的大量数据的运用主要是在云端处理，也即是说我们将传感器所产生的数据全部传输至云，然后进行计算，分析等后续处理。如此将产生大量的无效数据占用传输与计算资源的问题，所以此时如果能够通过AI技术达到在前端收集时进行预处理，就只用传输及计算信息中的关键帧。但随之而来的问题是如何在类似传感器的低端处理器上运载庞大而复杂的AI模型。因此这时模型压缩的课题就随之产生了，要把人工智能通过深度学习而训练产生的模型压缩到足以在数据搜集端运载以对数据进行初步的计算与分析。而这一课题就是唐老师为我们带来的报告的主要内容。

如何做到良好的模型压缩？唐老师将主流的做法进行了归类，分别是：Weight pruning 剪枝，Quantization 量化，以及Knowledge distillation 知识蒸馏。其中我对剪枝这一块的印象较为深刻，可能与自己曾经阅读过有关决策树相关的资料有关。在生成一棵决策树的过程中，很有可能为了最大可能的正确分类样本数据而导致产生结点过多而出现过度拟合的现象。这时就需要清除掉那些不能够给整棵树的决策能力带来泛化性能提升的结点。而在神经网络之中，剪枝的目的是为了减小神经网络参数数量，进而降低计算量。通过数据训练的大型神经网络经常因为应用场景的局限而存在一定的冗余，因此要做的就是在精确度没有不可接受的损失的前提下，尽可能的减少神经元的连接。唐老师在这一部分主要介绍了Non-structure pruning和Structured pruning的应用。Non-structure pruning通过压缩模型得到稀疏矩阵从而减小CPU的占用继而提高运送效率。然而一个很大的问题是，这种方法所能带来的速度提升与剪枝程度不成比例，若为了将运行速率提高到符合要求的水平，去掉的神经元连接过多可能会导致精度下降超过可接受的范围。于是在这时就引入了后者进行优化。量化的研究成果中，老师似乎（在这方面我的知识积累有些欠缺）主要提到了BNN 二值神经网络，通过将大部分的计算转化为计算机最为擅长的位运算实现而提高运算效率，并且通过二值化，能够将原本的数据如浮点数，只需用1 bit就可储存，减小了内存占用与开支能耗。在讲解知识蒸馏过程中，老师向我们展示了通过teacher-network带动student-network的例子。而其中最让我惊讶的是有时通过此法进行模型压缩之后，精确性不仅不会大幅度的下降，甚至还有精度提高的可能性。

随后唐老师介绍了他们目前正在攻关的自动结构化压缩研究方向以及滴滴公司这些年在人工智能以及无人驾驶方面的科研成果。其中给我印象较为深刻的一点是从学术界转身步入企业界的过程中，唐老师发现了科研领域与技术应用领域存在的不同。很多在实验室中的研究并不具备很大的落地可能，因为这些研究可能并没有考虑到成本因素以及现实场景中可能出现的突发情况。这让我们进一步明确技术应用需要理解其中原理，而计算机科学方面的研究人员也同样应当考虑到实际应用层面的困难与实现方法。最后唐老师对数学和英语的掌握也是令我们极为敬佩，也树立了更为清晰的奋斗方向与目标。

计算机科学与技术学院19级04班 王翰辉