智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

城市大脑的痛点与对策

高文

引用本文:
高文. 城市大脑的痛点与对策[J]. 智能系统学报, 2020, 15(4): 818–824.
GAO Wen. City brain: challenges and solution[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(4): 818–824.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202011038

城市大脑的痛点与对策

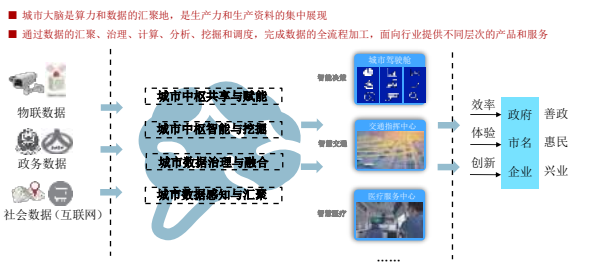
City brain: challenges and solution

高文

（北京大学 信息科学技术学院，北京 100871）

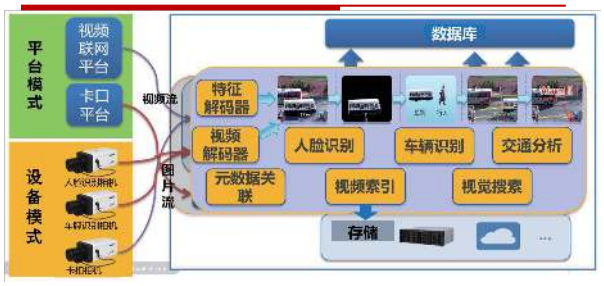
智慧城市是一个由传感器网络、云中心以及决策支持系统等要素组成的复杂系统，而城市大脑是智慧城市的核心，它将数据、算力、算法汇聚在一起，提供信息社会最强的生产力和生产资料。我们把互联网数据、政务数据和社会数据汇聚在一起，通过智能、数据、业务等中台服务打造城市大脑，结合云计算服务，就可以根据特定场景形成重大决策，支撑各种应用，提供便民服务，提升政府效率，提速企业创新。

城市大脑是智慧城市的核心



在智慧城市汇集的各种数据中，80%~90% 与图像、视频相关联。对城市大脑而言，“如何处理好海量图像和视频数据”极其关键。

城市大脑的核心是视觉认知计算(VCC)



现有的网络视觉感知系统有两种典型的应用模式：

1) 第 1 种模式是视频采集终端。摄像头是一个简单的传感与编码压缩装置，捕捉到图像或者视频后，进行编码压缩，然后传送至云端。云端可以存储，也可以将它读出解码，然后抽取特征进行分析识别，分类识别出人脸、车辆等目标或者聚集打架、车辆闯红灯等行为。

2) 第 2 种模式是智能终端。智能终端设备具有识别能力，在前端就把人脸或者车牌等信息识别出来，识别出来的结构化信息被传送到云端，直接可以分析使用。

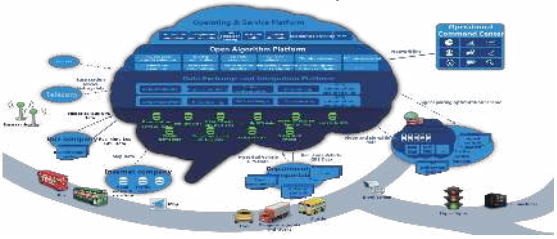
这两种模式各自都存在一些问题。如果仅仅使用视频采集终端，则传送回云端的数据是非结构化的，无法直接使用。若想分析使用这些视频数据，除了解码外，还要进行特征提取等工作，这需要在云端进行大量的计算，非常耗费算力资源。例如当传感器网络规模达到百万路摄像头时，可能需要超过百亿元规模的云计算服务器投入，即使真有这么多钱买服务器，其每年电力消耗也是一笔巨大的开销。如果全都使用智能终端，由于各终端厂家以及软件系统商使用的特征以及算法不统一，当原来系统中存在未被定义的物体分类识别以及行为分析时，不同厂商的智能设备互操作难度大，无法开展异构系统的新业务布局。所以，我们需要一个更好的系统，不仅云上算力资源配备需求不应过大，而且可以容易升级部署新的分析识别任务。

1 城市大脑 1.0

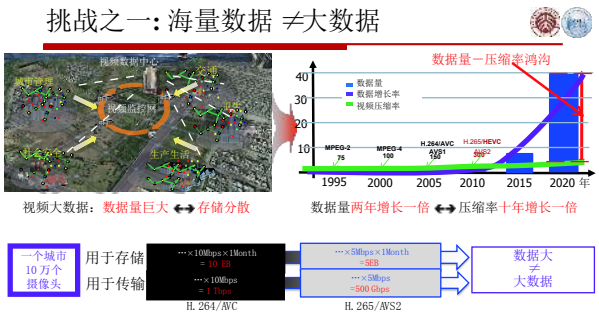
我们将仅仅由传感器网络和云中心两级组织架构组成的智慧城市系统，称为城市大脑 1.0。

城市大脑1.0—超大规模人工视觉系统应用

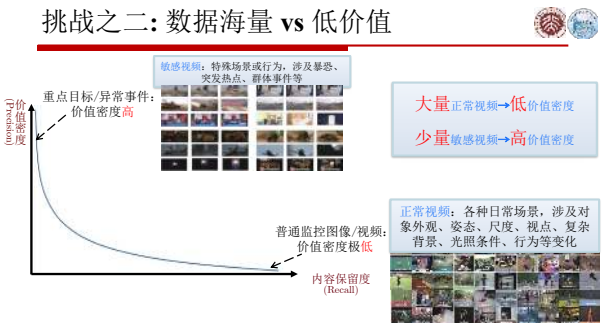
■ A computer vision system in cloud, connected with one or more camera network systems



城市大脑 1.0，虽然拥有海量数据，但是它并不等于大数据。因为 90% 的海量数据没有结构化，只是进行编码压缩后存储了起来。所以虽然数据是海量级别的，但这并不是大数据。



其次，这些非结构化视频数据除了作为案件的事后追踪可以发挥一定作用，事实上很难使用大数据挖掘工具找到有用的规律，因此价值比较低。这也是很多智慧城市的视频数据一段时间后（最短 2 周，最长 3 个月）就会被覆盖的重要原因。



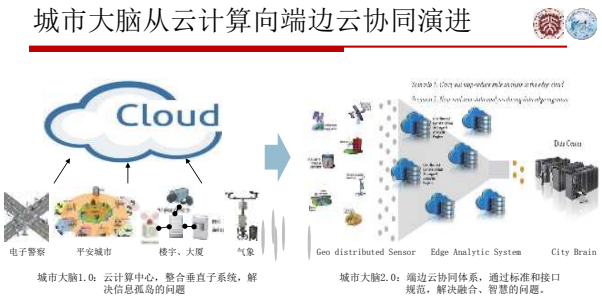
为什么会出现海量数据却是低价值的情况呢？问题的实质就是现有的城市大脑里的视频数据表达不到位，是非结构化的，即使有些声称已经结构化了也只是特定厂商针对自己的使用做了局部的结构化，没有形成真正可以开放给任何应用软件开发商使用的结构化数据。

要想解决上述问题，我们需要一种泛化能力更强的数据表达：基元表达，或者特征元数据表达。这些基元，既可以完成现有的任务，也可以完成现在还没有定义的任务。这个问题，在十年前还几乎是不可完成的任务，因为那时候手工特征盛行，识别分析系统的性能（准确度）都是与自己的手工特征紧密相关的，那是算法的核心竞争力。现在不同了，大家都是在使用深度神经网络，独具特点的手工特征已经不再当作核心机密不和外面交流了，深度网络特征已经成为首选。当然，即使是深度特征，数据表达想要得到一个比较好的结果，基于大数据的模型训练是必不可少的，那样系统整体才能做得更好。所以城市大脑应该有一套评测基准，包括系统的智力、性能（响应时间、并发、吞吐）、效率（耗电多大）等等。

城市大脑 1.0 的弊端在于实现智能的代价比较高，造价和耗电都非常惊人。如果希望城市大脑变得更智能，更高效，那就需要城市大脑 1.0 升级进化到 2.0，即边端云结合城市大脑。

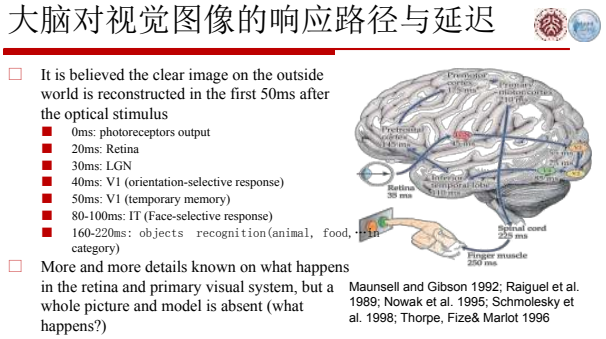
2　城市大脑 2.0

城市大脑 2.0 的关键就是任务合理划分：把原来的传感器网络与云中心一体化的系统架构，演变成边端云协同的系统架构。云上只需配备最低的算力，一部分计算放置于边缘，一部分计算分配给终端，这样组合起来使得整个系统最优。



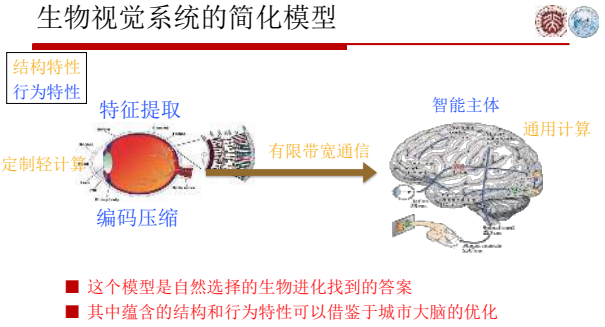
系统升级策略可以借鉴人的视觉系统。人的视觉系统是一个非常合理、能效比非常高的系统。人消耗的能量，相当于 20 W 电灯泡的能耗，但是我们的视觉系统比任何超级计算机构成的计算机视觉系统分析和识别能力都不差，有时还更好。人的视觉系统为何可以做到如此低功耗、高效率？人的视觉系统主要由 3 部分组成：眼睛、视觉通路和大脑的视觉野。3 个部分分工严密，比如来自眼睛视网膜的信号，通过视觉通路传到大脑不同的视觉野，不同的视觉野做出不同的响应，就可以完成诸如感知、识别、决策等很多任

务。不同的感知路径或者不同复杂度的任务，其响应时间是不一样的。下图是 1992 年的一张研究成果示意图。当给一个人下达指令:“你给我按一下绿色按钮”，这个指令的执行是经过一定延迟的，首先视网膜有 35 ms 的延迟，从视网膜到下一个环节又有 30 ms 的延迟，最后到肌肉带动手指执行按下按钮的动作，大概有 250 ms 的延迟。这个例子告诉我们，对于不同的任务，我们整个视觉通道和大脑的处理分工是非常严密的，简单的任务响应快，复杂的任务响应慢。只有分工合作，系统才能做到能量最优化。



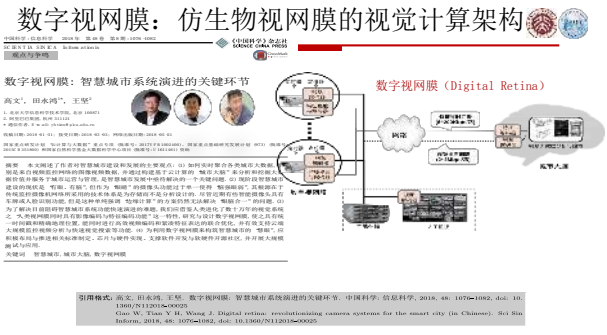
视觉系统最前端是视网膜，视网膜由感光细胞、双极细胞和神经节细胞 3 类细胞组成。视网膜大概有 1.2 亿~1.26 亿个感光细胞，其中有锥状细胞和杆状细胞，锥状细胞有 600 多万个，杆状细胞有 1.2 亿个，它们可以感知光线的强弱，这些感光细胞的输出信号通过双极细胞，最后汇聚到神经节细胞，进入神经纤维、视觉通道，并传输至大脑。神经节细胞的数量只有差不多 100 万个，也就是说，从视网膜到视神经，已经有大约 125:1 的缩减，这个减缩可以理解成视觉信号的压缩，或者特征压缩，该压缩过程对整个大脑的有效工作起到非常关键的作用。当然这不仅仅是压缩处理，而是特征编码，与后续的感知紧密相关。根据任务的复杂程度不同，所需提取的视觉特征也不同；简单的任务就会优先采取快速处理和响应的策略，复杂的任务就把相关的信息往后传。

一个生物识别系统的简化模型：感知信息从视网膜到大脑，经过特征提取和编码压缩后向后传输，最终传到智能主体 (脑)。因此，在视网膜端，完成的是定制化轻量级计算，然后通过视神经这样一个有限带宽的通信通路将视网膜计算结果送到智能主体。仿生视网膜的架构具有非常好的能量优化特点，为了使整个城市大脑达到能量优化或者能量高效化，就可以按照仿生视网膜的架构来构建城市大脑。



所以信息从视网膜传到大脑，是经过特征压缩处理的，又称为特征编码，和现在传统的图像编码不同的是，它是把特征编码压缩的结果送到大脑中去。

另外，现在城市大脑不能仅传特征，也要传压缩图像，因为某些情况下还需要用人眼确认，这就使得城市大脑的架构和人类的视觉系统并不完全一样，而是两个综合或者绑定的系统。基于以上分析，即可进行城市大脑 2.0 的设计。显然它必须是一个边、端、云合理分工的系统，在这个系统中，边、端、云结合的最核心的技术叫做数字视网膜，是整个城市大脑 2.0 的基本架构，我们把它叫做仿生视网膜的计算架构。



3　数字视网膜

我们给数字视网膜定义了 8 个特征属性，原则上分成三大组。

第 1 组特征属性是与时空有关的。一个数字视网膜的终端必须要有全局统一的时空 ID，包括全网统一的时间戳和精确的地理位置，比如 GPS 或者北斗提供的位置，从而便于城市大脑的同步和标校。

第 2 组特征属性简单来说是视频编码+特征编码+联合优化，这是所有摄像头都应该支持的一项功能属性。而当前绝大部分摄像头只支持

视频编码，没有特征编码。视频编码是为了存储和离线观看影像重构。特征编码是为了模式识别和场景理解的紧凑特征表达。由于城市大脑 2.0 至少有两个码流，一个是视频编码压缩流，一个是特征编码压缩流，这两个码流会捆绑到一起进行传输，因此，还需要通过联合优化，把带宽合理分配给视频编码和特征编码，使得整个系统是最优的。

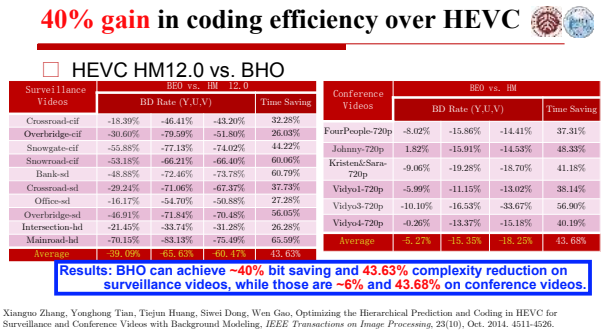
第 3 组特征属性，简单来说就是模型可更新、注意可调节、软件可定义。模型可更新是指当模型需要切换或升级时，终端要能够进行实时更新，以更好地支持多种神经网络和算法。注意可调整是指摄像头能够自动调节焦距、拍摄角度等配置参数。软件可定义则是指可以通过软件定义的方法对系统进行自动升级。如果具备这 3 个特点，终端就可以做得非常智能。

当然，要想把数字视网膜技术全部用起来，这里面有一些使能技术。

第一个使能技术是视频编码。目前城市大脑、监控系统都离不开视频编码，摄像头里面都有一个视频编码芯片，视频编码芯片使用的标准，最早期是 H.264 或者 AVS+，最近开始转变为 H.265 或者 AVS2 的标准，未来不久就会用上 AVS3 或者 AV1 或者 H.266，该标准几乎每 10 年就会更新一代，编码效率相应地提高一倍。

一段视频是一个图像序列，图像序列里包含了很多数据的冗余，基本上可分为三大类：一类是和空间有关的冗余，一类是和时间有关的冗余，另外一类是和编码有关的冗余。为了消除冗余数据，就要对视频进行编码压缩。现在整个视频编码用的算法一般是混合视频编码架构，即将上述 3 种主流冗余数据用不同的算法去除掉。比如为了去除空间冗余，一般采用正交变换 (DCT 变换等)；为了去除时间冗余，就是帧与帧之间的冗余，一般会采取预测编码，比如各种各样的滤波器；为了使编码的分配最符合熵的定义，我们使用信息熵编码来去除编码上的冗余。这 3 类冗余都去除了，整个视频流就可以压缩得很小。

要把视频编码做好，算法要做得很精，随着时间的推移，可以用计算、带宽把这些冗余一点点都去除掉。当然，这些年我们除了不断地优化算法之外，还提出了一种背景建模技术，使得编码效率在原有的技术上再提高一倍。



Xianguo Zhang, Yonghong Tian, Tiejun Huang, Siwei Dong, Wen Gao, Optimizing the Hierarchical Prediction and Coding in HEVC for Surveillance and Conference Videos with Background Modeling, *IEEE Transactions on Image Processing*, 23(10), Oct. 2014. 4511-4526.

这里有很详细的一些数据测试作为依据，而且相关研究成果都已经发表论文，比如 2014 年我们在 IEEE T-IP 发表了一篇论文，里面有这样一些研究结果。AVS2 于 2016 年成为我国的视频编码标准，同时它也是 IEEE1857 标准的第 4 部分。目前正在制定的 AVS3，是 IEEE 1857 标准的第 10 部分。2019 年 3 月发布了 AVS3 标准第 1 版，而 H.266 第 1 版直到 2020 年 7 月才发布，我们超前了 H.266 一年零三个月，这是有史以来第一次，国内标准超前于国际标准完成。AVS3 标准 2019 年 3 月第 1 版发布以后，同年 9 月海思就完成了芯片制造，这款芯片在阿姆斯特丹的一次广电展上一经面市，就引起了很大的轰动。它可以支持 AVS3、8K 解码，支持高动态和每秒 120 帧速率。该款芯片现在已经装配于很多 4K 电视、8K 电视、机顶盒等。

第 2 个使能技术就是特征编码，是非常关键的一个使能技术，该技术的标准有两部分核心内容，一部分叫 CDVS，另一部分叫 CDVA，现在也是国际标准 MPEG-7 里的两个部分，一个是 MPEG-7 第 13 部分，2015 年 9 月发布，一个是 MPEG-7 第 15 部分，2019 年 7 月分布。

从图像中提取出来的特征数据可能很大，如果不压缩的话，特征数据很可能比图像本身都大，因此，同样需要对视觉特征进行编码压缩。

如何进行特征压缩也是一个值得考虑的问题。一种途径是先把图像降质编码传过去，然后提取特征，再进行识别；另一种途径是先把特征提取出来，然后把特征传过去再识别。这两种途径存在一个剪刀差，可能导致识别率相差百分之二十、三十甚至更高。因为先进行图像压缩可能造成一些有用特征的丢失，传统编码压缩，倾向保留符合人眼视觉特性的公共部分、压缩掉一些非公共的、非常见的信息，而非常见的部分恰恰可能是面向机器识别的有用特征，所以该压缩处理很可能导致识别率的下降，因

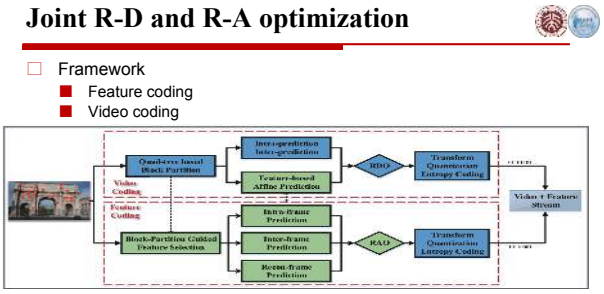
此我们采取先提取特征然后在云端识别的技术策略，就可以保证特征信息不被视频编码流程所丢失。

先提特征，怎样使提取的特征体量比较小？我们初期针对手工特征，设计了低比特、高性能、低复杂度的全局与局部特征表示，形成对图像的一种全局紧凑描述，实现图像快速比对，并支持局部特征的快速匹配，检测几何一致性。我们做了第 1 版以后，又专门做了一个面向深度学习特征的编码压缩框架，主要是针对小视频来做的，提供了数据驱动的几何与语义不变性紧凑特征表示。有了这 2 个部分以后，基本上可以应对图像特征编码和视频特征编码这两个需求。

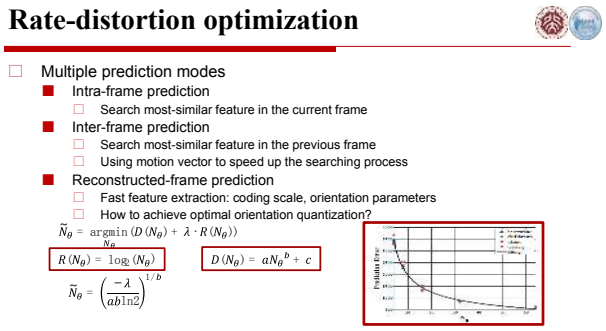
图像特征编码即 CDVS，视频特征编码即 CDVA。CDVS 是手工特征，使用的是一个类 SIFT 的特征集，当输入的比特数据比较少时，就给出一些比较宏观的特征。CDVS 为单幅图像提供了 512 B/1 KB/2 KB/4 KB/8 KB/16 KB 的可伸缩码率，这一特性有利于克服无线传输面临的带宽受限、带宽波动等技术挑战。基于这样的思路，用这种类 SIFT，我们提出了一个特征紧凑表达的标准，然后评测它的性能，经过几年的时间，性能越来越高，最后固定下来。对比图像压缩，特征压缩效率提升百倍 (测算依据：压缩图像大小 400 KB/典型特征大小 4 KB)。CDVS 从 2012 年 2 月份启动，到 2015 年 6 月份完成，最后成为国际标准，投入了将近 4 年的时间。CDVS 完成后，标准化组织团队便立即转向研究利用深度学习进行视频分析特征压缩的问题，花费两三年的时间完成了技术攻关，可以利用深度网络对短视频进行特征提取与表达，并且特征的性能一直在逐步提高，在不同网络环境下，其特征提取和特征识别的效率也在逐步提高，并进一步实现了融合深度学习特征与传统手工特征的高性能视频特征压缩技术。对比源端视频，实现了近万倍压缩比 (测算依据：源端视频流码率 1.5 Gbps/特征流码率 150 Kbps)。现在 CDVA 也已经成为国际标准。

第 3 个使能技术叫做联合优化。所谓联合优化，就是在视频编码和特征编码之间，找到一个最优的结合点，使得这两个流捆绑到一起的时候，码率分配是最优的，这样送到云里，它们合起来是最优的。怎么能够做到最优呢？因为各自的优化模型都是有的。

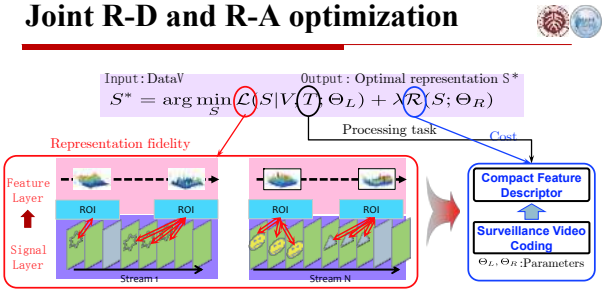
比如我们看到的上面这部分，是一个视频编码优化的流程，上面的虚线框是视频编码，下面的实线是特征编码，这两个编码合成一个流，就是视频和特征流。



将视频和特征流放入联合优化流程中一起优化。视频编码的优化模型叫 RDO，RDO 就是给定码率条件下损失最小的优化模型，它的优化曲线就是右下角这个曲线。在识别特征表达这一块，有一个 RAO，就是给定码率条件下，让精确度最高的优化模型。



这个优化模型给的曲线是反过来的，所以把这两个需要优化的东西放到一个优化函数里面表达出来，就是这张图的表达，据此联合求解一个优化的解，这就是联合优化。



第 4 个使能技术是深度学习模型编码的使能技术，使用多个源模型来增强目标模型学习，提升目标模型的性能与泛化性，并通过构建模型之间的预测机制，实现增量式模型更新，降低模型更新带来的码率开销，提升模型部署效率。

多模型重用既包括现有模型的重复使用，也

包括根据目标模型训练所得到的优化模型部署使用。如果结合边端云计算框架，在学习体系中用好多模型重用，那么性能就可以得到大幅提升。因此，如何在多模型重用过程中，使用模型编码快速地更新模型，促使性能不断提升，就是模型编码的主要动机，这样就可以在模型训练完成、压缩好后快速推到终端去升级模型。

上述使能技术最终可以在芯片里实现，这类芯片称为数字视网膜芯片，目前北京大学杭州研究院的一个下属公司已经完成了芯片的设计制造，芯片型号是 GV9531，支持上文所述的数字视网膜 3 组 8 个特性。并且研发了基于该芯片的板卡，比如 4 颗芯片的卡、16 颗芯片的卡，这些板卡已经可以支持边缘端，支持上百路甚至几百路的摄像头数字视网膜特征提取的传输。

除了数字视网膜本身以外，配合人工智能技术的发展，当前也在推动中国的一些 AI 技术成为国家标准，包括神经网络模型表示与压缩的标准、城市级大数据汇集关联的规范和标准，同时也在规划这些标准研究制定的路线图及时间表等。

数字视网膜简单来说是 3 个编码流合并的系统，即视频流、特征流和模型流，其中，视频流和特征流是最主要的部分，而模型流只是在需要更新模型时将模型编码压缩后从云端推到边缘或者终端上，进行一些增量更新。

有了数字视网膜，就可以使得城市大脑边缘或者终端的效能比更高，从而减少云端的算力，同时使云端的响应更精确、更快速。

为了配合这个工作，目前在鹏城实验室有比较完整的设计和规划，包括一些中台、业务支撑以及应用等系统。我们把城市大脑 2.0 的数字视网膜简称为云脑视网膜，然后利用鹏城云脑的算力去提升它的能力。到目前为止，鹏城云脑的建设已经投入了几十亿元，拥有了 100 P 的算力是目前国内算力最大的一套 AI 训练系统。鹏城云脑仍处于建设阶段，未来将会成为更强的系统。我们目前已经研发了一套数字视网膜原型系统，支持数据采集、上传、标注、训练，支持采用基于数字视网膜芯片的终端、服务器进行提取，然后进一步分析和识别。

该原型系统已经开始汇聚越来越多的数据，技术也越来越成熟，包括了大数据，人工智能等开放平台，系统中运行着各种各样的与硬件相互配合的参考软件，最上层则是开源算法训练，以此为基础，将来鹏城云脑会对城市大脑进行更强有力的支撑。

目前已有一些演示验证案例，例如，利用深圳交警提供的数据进行系统验证、视频追踪等等。同时，在深圳市光明区若干路段也开展了一些现场测试和示范应用，验证结果表明，系统对于停车、拥堵等事件都可以很好地分析和发现。

4　结束语

上面是城市大脑 2.0 到现在为止的一些进展情况。城市大脑 1.0 是一个以云计算为核心的系统，由于系统分工协调不好，所以成本比较高，响应速度慢，数据的可利用度也比较低。城市大脑 2.0 借鉴人类的视觉系统，提出了一个性能更优异的体系架构，该体系架构需要数字视网膜的思路、技术及其标准化等工作的支撑与配合，目前，相关思路、技术、标准化都已逐步到位。

数字视网膜系统，可以使现有城市大脑 1.0 在编码方面节省 50% 的存储和带宽，在云资源耗费上节省 90% 以上的计算算力，而且对图像特征的提取和分析延迟更低、精度更高，这是数字视网膜带给城市大脑 2.0 的一个好处。当然，数字视网膜系统的完善还需要一段时间，还需要在更多的应用中进行验证，当相关技术成熟、标准制定完成时，城市大脑 2.0 真正运营起来，就会对中国的城市化、智能城市发展等方面发挥较大

的贡献。

致谢：本文根据我在 2020 年 CCF-GAIR 上的大会报告录音整理而成的，文稿得到过田永鸿教授、马思伟教授、段凌宇教授、贾惠柱副研究员、张伟先生等同事的修改。在此一并对他们表示感谢。

参考文献：

[1] 高文, 田永鸿, 王坚. 数字视网膜: 智慧城市系统演进的关键环节 [J]. 中国科学: 信息科学, 2018, 48(8): 1076–1082.

[2] ZHANG Xianguo, HUANG Tiejun, TIAN Yonghong, et al. Background-modeling based adaptive prediction for surveillance video coding[J]. IEEE transactions on image processing, 2014, 23(2): 769–784.

[3] 高文, 等. 信息技术: 智能媒体编码 (第 2 部分: 视频), 新一代人工智能产业技术创新战略联盟团体标准 [S]. 2019.

[4] BROSS B, CHEN J, LIU S, et al. Versatile Video Coding, ITU-T and ISO/IEC JVET-S2001, 2020.

[5] DUAN Lingyu, CHANDRASEKHAR Vijay, CHEN Jie, et al. Overview of the MPEG-CDVS standard[J]. IEEE transactions on image processing, 2016, 25(1): 179–194.

[6] DUAN Lingyu, CHANDRASEKHAR Vijay, WANG Shiqi, et al. Compact descriptors for video analysis: the emerging MPEG standard[J]. IEEE multimedia, 2019, 26(2): 44–54.

[7] ZHANG Xiang, MA Siwei, WANG Shiqi, et al. A joint compression scheme of video feature descriptors and visual content[J]. IEEE transactions on image processing, 2017, 26(2): 633–647.

作者简介：

高文，中国工程院院士、北京大学博雅讲席教授，鹏城实验室主任，新一代人工智能产业技术创新战略联盟理事长，全国信息技术标准化委员会副主任，数字音视频编解码技术标准 (AVS) 工作组组长，国际电气和电子工程师协会会士 (IEEE Fellow)、美国计算机协会会士 (ACM Fellow)。主要从事人工智能应用和多媒体技术、计算机视觉、模式识别与图像处理、虚拟现实方面的研究，主要著作有《数字视频编码技术原理》《Advanced Video Coding Systems》等。在本领域国际期刊上发表学术论文 200 余篇，国际会议论文 700 余篇。

中文引用格式：高文. 城市大脑的痛点与对策 [J]. 智能系统学报, 2020, 15(4): 818–824.

英文引用格式：GAO Wen. City brain: challenges and solution[J]. CAAI transactions on intelligent systems, 2020, 15(4): 818–824.