语义分割综述

李想北京理工大学徐特立学院30081702班，学号1120171688，北京理工大学

(1120171688@bit.edu.cn)

摘要 由本文为北京理工大学计算机体系结构小组报告，由组长李想和组员张翰澄完成。本文聚焦时下正流行的深度学习，并以深度学习芯片为切入点，调查和分析了多种深度学习芯片的结构和其背后的设计思想，从而形成了对于深度学习相关芯片例如TPU，寒武纪DianNao的一些理解。另一方面，本文虽然从深度学习芯片的结构出发，但是作为北京理工大学的学生同时作为计算机体系结构课程报告，没有局限于机械的搜索和堆砌目前已有芯片的数据，而是充分结合自身软硬件编程经验，从计算机体系结构的高层次，对于以深度学习芯片为首的芯片发展前景作出了分析和理解。

关键词课程报告，深度学习，芯片，体系结构，分析理解，专用集成电路

深度学习已经能够超越传统的方法，在语义分割领域取得更好的表现，本文总结到2020年为止比较出色的方法，并且给出分类。

## Inruduction

深度神经网络DNN是深度学习领域最重要的部分，因此现在很多芯片开始针对于DNN的运算方式进行特殊的设计来提高运算速度。DNN中常用的的网络层：卷积层和全联接层的核心就是乘加运算(MAC)，这种运算可以通过并行计算的方法有效地提升效率。在体系结构课程中我们学过，开发并行性的途

* 1. **XXXX**

CPU和GPU通过SIMD，SIMT的方式来并行MAC计算。通常来说，在这些平台上，常用的神经网络如卷积层和全联接层会被表示成矩阵乘法。如今CPU有很多软件库如Open-BLAS，Intel MKL；GPU有cuBLAS，cuDNN等等来优化矩阵乘法。这些平台上的矩阵乘法可以通过对数据进行计算转换以减少乘法次数来进一步加快，同时仍然给出相同的按位结果。 通常，这样做的代价是增加的添加数量和更不规则的数据访问模式。

* 1. **XXXX**

对于DNN来说，计算性能的瓶颈在于访存。每一次MAC需要三次读内存操作，分别读取：权重，输入数据，之前计算的部分和结果，需要一次写操作来更新计算的部分和。通常来说这些访存操作都是访问片外DRAM，也就是说大量的访存会严重影响计算速度和能耗。

而一些基于空间重叠技术的加速器通过设置片上的大缓冲区来解决这样的一个瓶颈。从已有的一些加速器中我们可以观察到，现有的加速器更倾向于把缓冲区分为三个部分，分别存储权重，输入和输出，这样可以减少它们之间的干扰，提高效率。

## 2．DL-based models

## 2.1．Encoder-Decoder based model

Badrinarayanan提出了SegNet[1]，This core trainable segmentation engine consists of an encoder network, a corresponding decoder network followed by a pixel-wise classification layer (figure1). specifically, the decoder uses pooling indices computed in the max-pooling step of the corresponding encoder to perform non-linear upsampling. This eliminates the need for learning to upsample. The upsampled maps are sparse and are then convolved with trainable filters to produce dense feature maps. The author compared the archetechure with FCN and DeepLab series which reveal the memory versus accuracy trade-off for achieving good segmentation results.

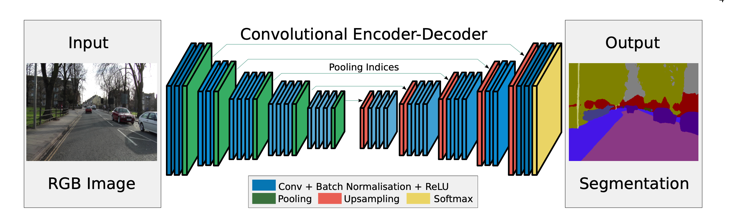
****

Figure1 SegNet has no fully-connected layers; hence, the model is fully convolutional. A decoder up-samples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). From[1]

There are also encoder-decoder model for medical and biomedical image segmentation. Medical and biomedical segmentation is faced with two challenges 1) there are very little training data and it’s important to let the network learn the invariance to such deformation without seeing the transformation in the annotated image corpus because deformation is common in tissues. 2)separation of touching objects. Therefore Ronneberger et al. [2] proposed the U-Net for segmenting biological microscopy images. The network architecture is illustrated in figure 2, it down-sample the feature maps like many architecture do. Feature maps from the down-sample part are copied to the up-sample part to keep the low-level information. For data argumentation, Random elastic deformations of the training samples seem to be the key concept to train a segmentation network with very few annotated images. they generate smooth deformations using random displacement vectors on a coarse 3 by 3 grid. The displacements are sampled from a Gaussian distribution with 10 pixels standard deviation. Per-pixel displacements are then computed using bicubic interpolation. Drop-out layers at the end of the contracting path perform further implicit data augmentation.

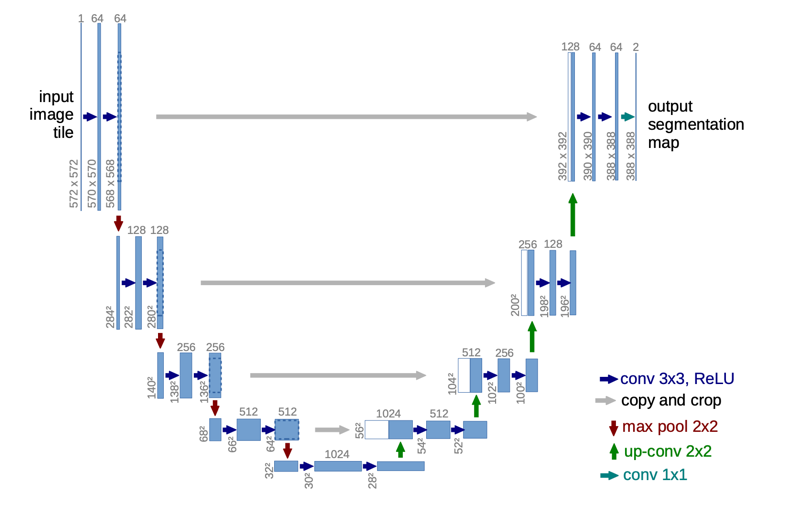
****

Figure2 U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations from [2].

There are also extensions of U-Net for different kind of images such as …

V-Net is another popular model for 3D image segmentation proposed by F. Milletari[3], they introduce a novel object function based on Dice coefficient which can deal with the situation where there is a strong imbalance between foregrounds and backgrounds. Also, they augment the data applying random non-linear transformations and histogram matching.

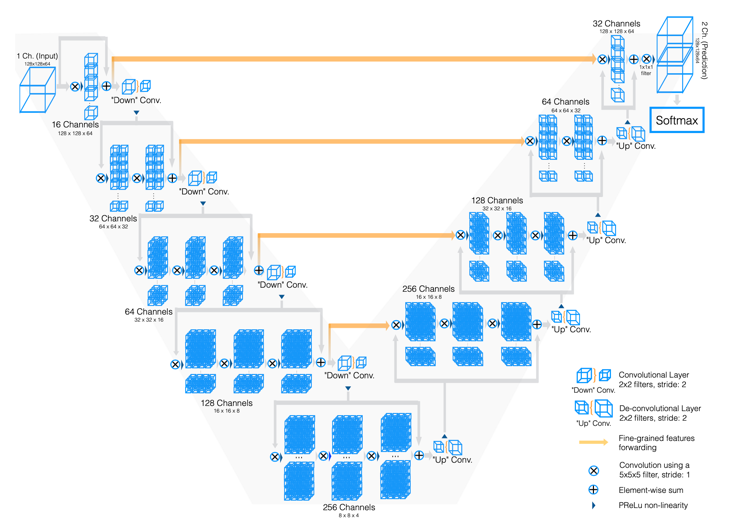


Figure3 Schematic representation of our network architecture.

## 2.2．Dilated Convolutional Models

Dilated convolution changes the dilation rate of the convolutional layers to have larger RF shown in figure4.

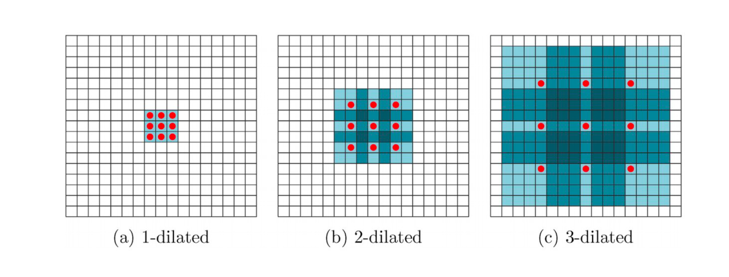


Figure4 Dilation convolution with different dilation rate

Many recent work uses this technique. Some most important include the DeepLab family[6], multi-scale context aggregation[8], dense upsampleing convolution and hybrid dilatedconvolution (DUC-HDC)[9], Dense connecte Atrous Spatial Pyramid Pooling(DenseASPP)[], and the efficient neural network(ENet)[].

DeepLabv2[6] mainly make three contribution that are experimentally shown to have substantial practical merits. First, it highlight the ‘atrous convolution’, as a powerful tool in dense prediction tasks. Atrous convolution allows us to explicitly control the resolution at which feature response are computed within the DCNN and also effectively enlarge the field of view to incorporate larger context without increasing the parameters. Second, it propose atrous spatial pyramid pooling(ASPP) to robustly segment objects at multiply scales. Third, it improve the localization of object boundaries by combining method from DCNNs and probabilistic graphical models. The commonly deployed combination of max-pooling and down-sampling in DCNNs achieves invariance but has a toll on localization accuracy. Chen overcome this by combining the responses at the final layer with a fully connected Conditional Random Field. Deeplab reaching 79.7% mIOU in the Pascal VOC-2012 test set.

Subsequently, *Chen, et al.* [7] proposed Deeplabv3, in this work, to handle the problem of segmenting objects at multiply scales, they design modules which emply atrous convolution in cascade or in parallel to capture muti-scale context by adopting multiply atrous rates.

Subsequently,

## 3． XXXXXX

**参 考 文 献**

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmenta- tion,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 12, pp. 2481–2495, 2017. [2] Norman P. Jouppi et al. In-Datacenter Performance Analysis of a Tensor Processing Unit, 44 ISCA, Toronto, Canada, June 26, 2017

[2] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networksfor biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[3] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmenta- tion,” in 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016, pp. 565–571. [4] Guilin Chen, Shen Ma, Yang Guo,

[5] Vivienne Sze, Yu-Hsin Chen et al, Efficient Processing of Deep Neural Networks: A Tutorial and Survey, Journal of Computer Research and Development, IEEE.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convo- lutional nets, atrous convolution, and fully connected crfs,” IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834–848, 2017.

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” arXiv preprint arXiv:1706.05587, 2017.

[8] F.Yu and V.Koltun, “Multi-scale context aggregation by dilated convolutions,” arXiv preprint arXiv:1511.07122, 2015.