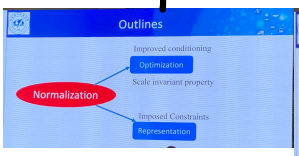
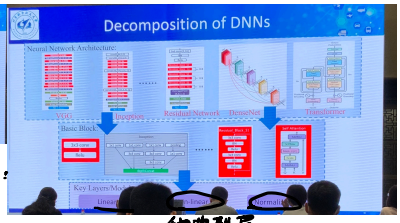
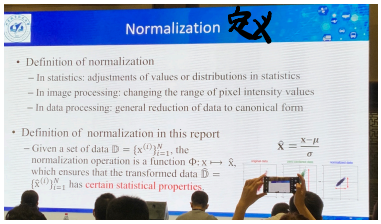
Workshop 3



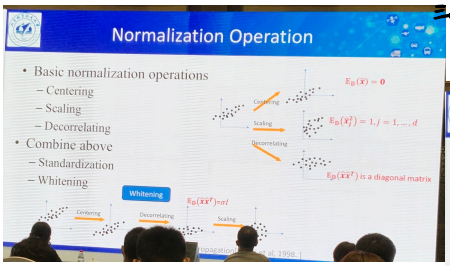


基本normalization方式

normalization出现原因：
gradient vanishing / explosion
activation vanishing / explosion → recurrent network.
网络判别能力差的开始
隐式调整方式：初始化
显式调整方式：normalization

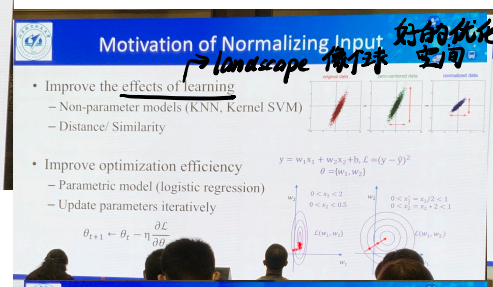


一种类型层

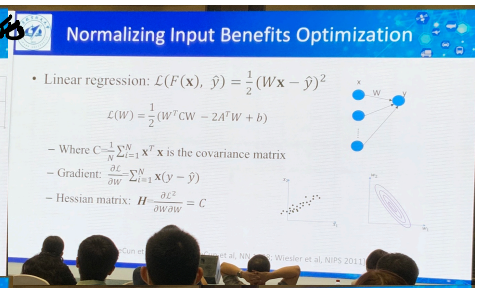


BN

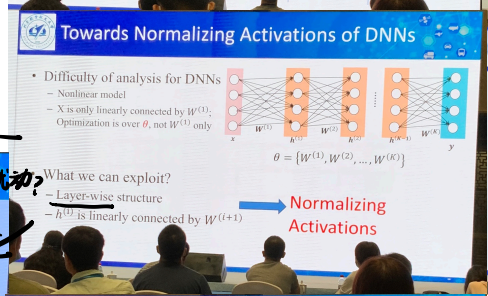
BN motivation

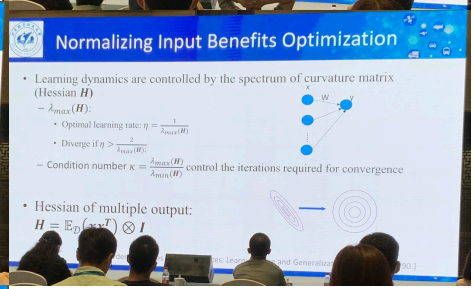


landscape 像个球 好的优化空间



线性模型

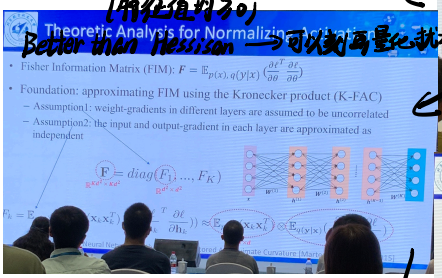


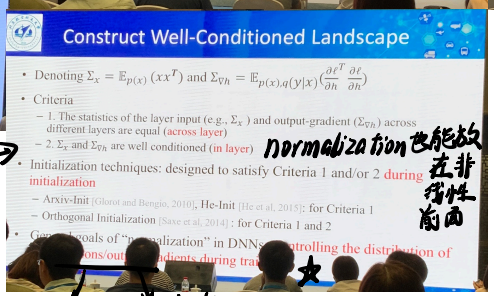


H=WX

(特征值为0)

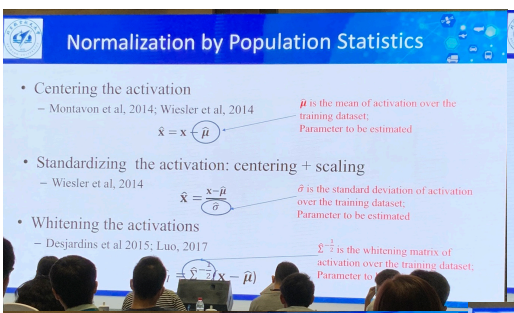
Better than Hessian → 可以数学上证明？

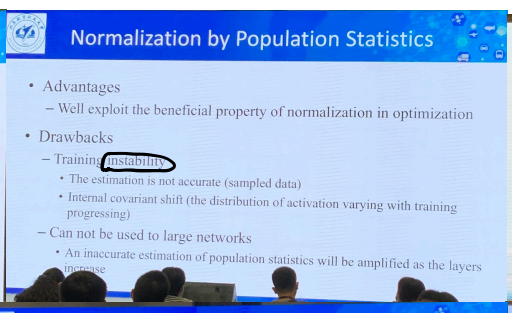


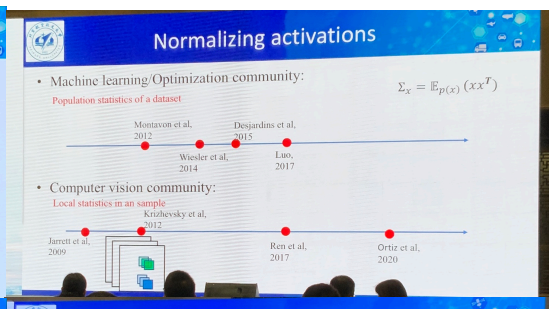


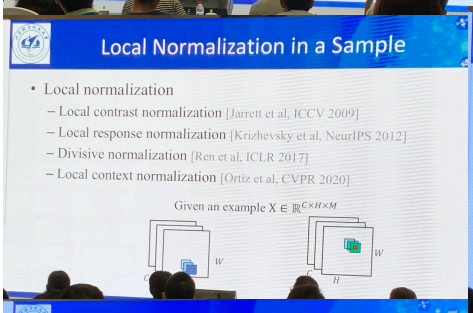
normalization也能放在非线性前面

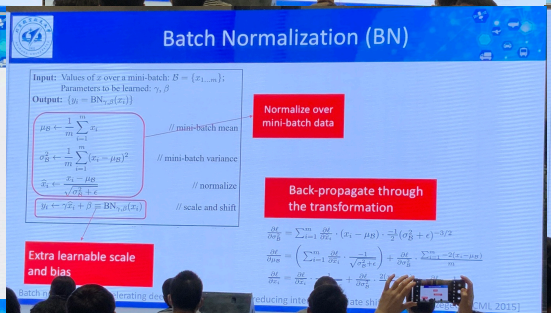
activation outputs

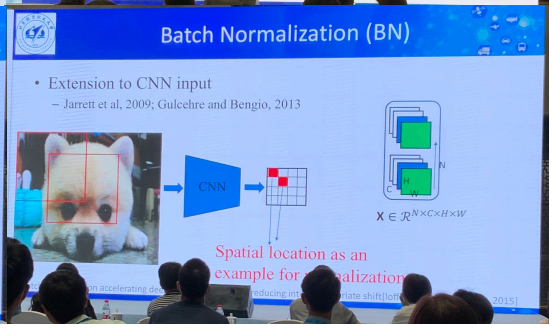


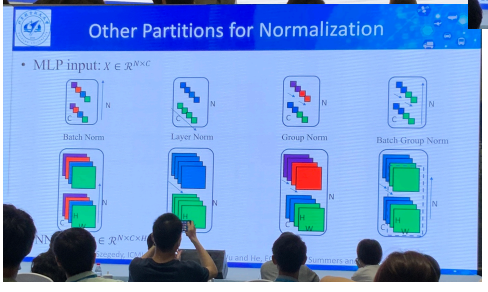




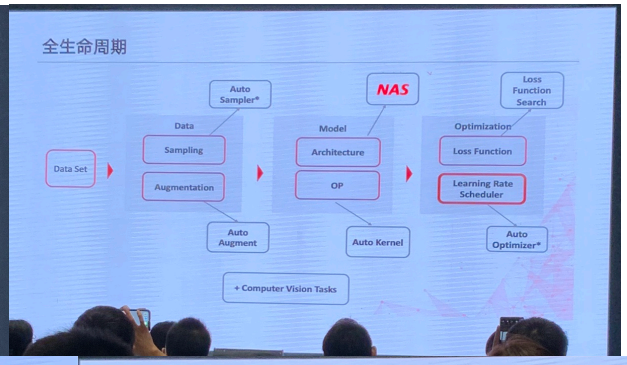


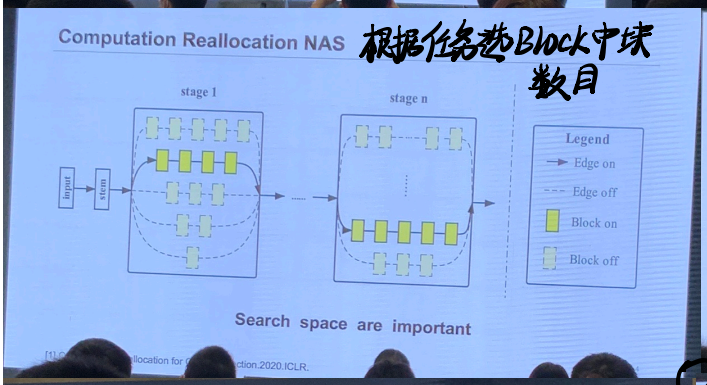


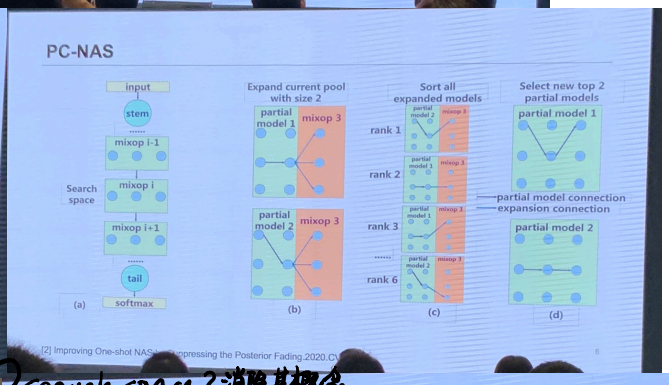


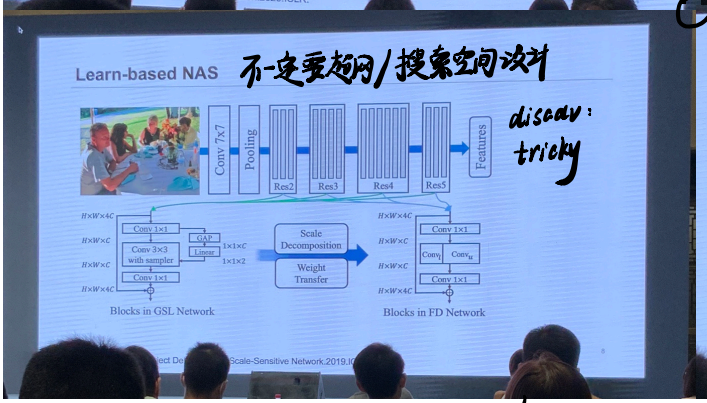


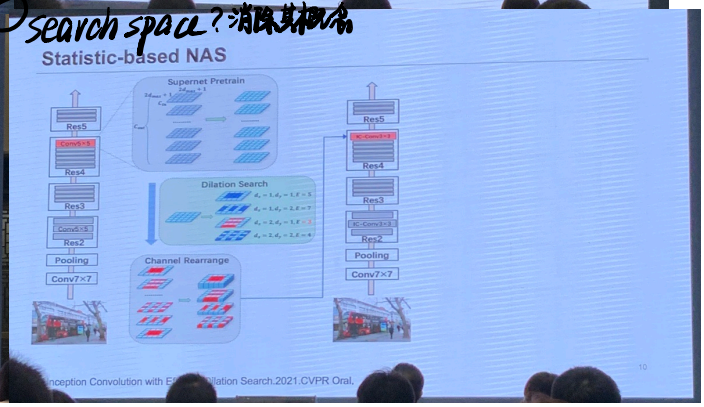
Workshop 4 全生命周期的自动化机器学习



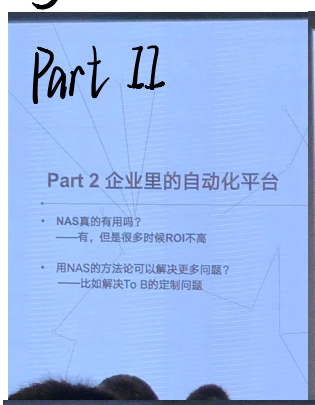


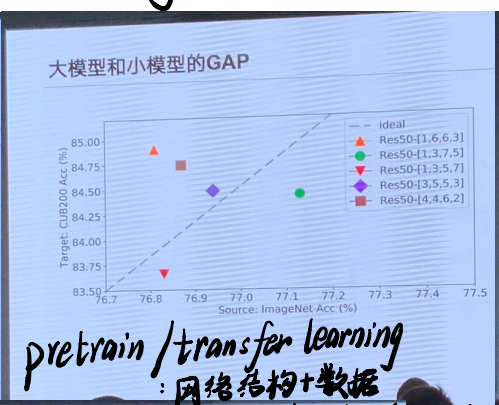


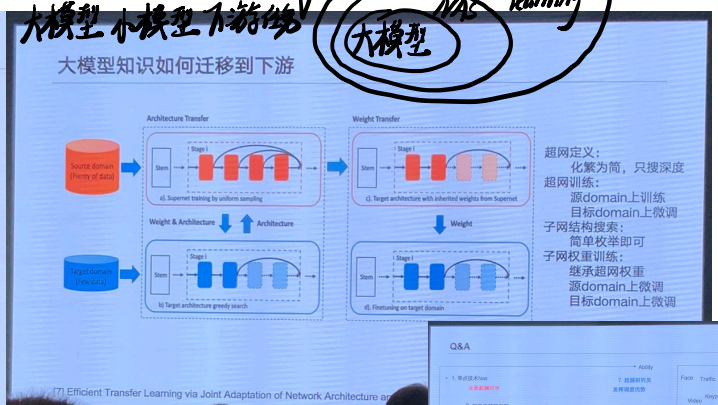




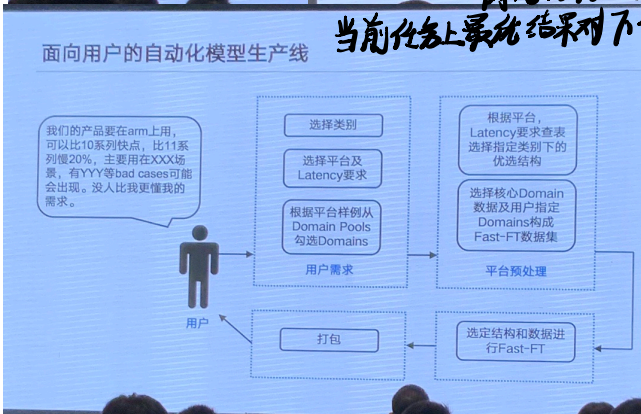
Beyond NAS { Auto sampling / AutoAug.

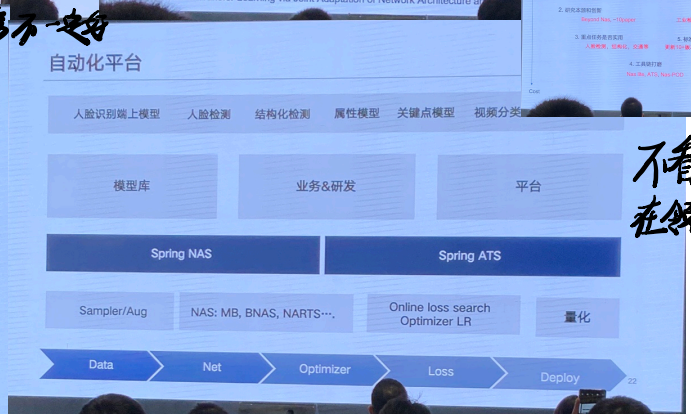


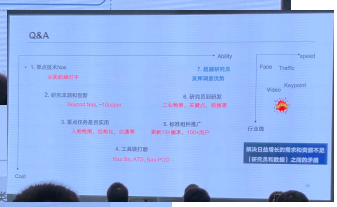




pretrain / transfer learning
：网络结构+数据
当前任务上最优结果对下个任务不一定好







不看好AutoML在领域自动应用

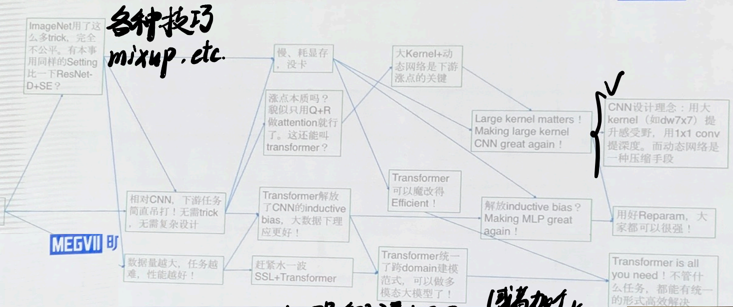
Workshop 5 transformer

近年研究热点回顾

- 层数更深、性能更强的架构
- 轻量级架构、高推理效率
- 自动化模型设计、神经网络架构搜索（NAS）
- 动态模型
- Attention Models、Vision Transformers (ViTs)

如何看待Vision Transformers？

MEGVII 旷视



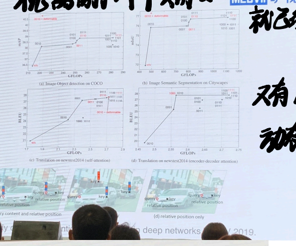
理解ViTs：潜在的优势

- 灵活的数据形式（Tensor、集合、序列、图）
- 长程关系建模能力
- 更强的表示能力 [1]
- 架构的合理性 [2]
- 对遮挡、噪声的稳健性 [3, 4]

[1] Cordonnier, Jean-Baptiste, Andreas Loukas, and Martin Jaggi. "On the relationship between self-attention and convolutional layers." ICLR 2020.
[2] Dong, Yihe, Jean-Baptiste Cordonnier, and Andreas Loukas. "…is not all you need: Pure attention loses rank doubly exponentially with depth."
[3] Xie, Cihang, et al. "Feature de… for improving adversarial robustness." CVPR 2019.
[4] Paul, Sayak, and Pin-Yu Chen…

Rethinking ViTs

- Multi-head self-attention (MHSA)可能不是必须的



机器翻译中用Q
(去掉加Q K 就已经很好)
有KQ最好？
动态机制？

Rethinking ViTs (cont'd)

MEGVII 旷视

- ViTs的设计要素
  - Sparse connectivity
  - Weight sharing
  - Dynamic weight

ViT设计要素
潜在优势并
非独有

→ large kernel design
high order relation modeling

[1] Zhao, Yucheng, et al. "A Battle of Network Structures: An Empirical Study of CNN, Transformer, and MLP."
[2] Han, Qi, et al. "Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight."

Advantages of Large Kernels

MEGVII 旷视

①
- 大卷积核可以更高效地提升有效感受野（ERF）

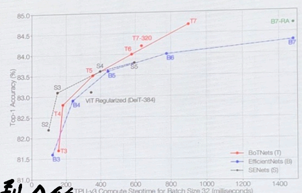
$$\sqrt{\mathrm{Var}[S_n]} = \sqrt{n}\sqrt{\sum_{m=0}^{k-1}\frac{m^2}{k} - \left(\sum_{m=0}^{k-1}\frac{m}{k}\right)^2} = \sqrt{\frac{n(k^2-1)}{12}} = O(k\sqrt{n})$$

Advantages of Large Kernels (cont'd)

MEGVII 旷视

②（层数减少）
- 大卷积核可以部分回避模型深度增加带来的优化难题
  - VGG-style models 难以做深 [1]
  - ResNets 的有效深度可能很浅 [2, 3]

6~10层？
→ BN可能背锅
限制能量
→ capacity 再强也看不到全貌



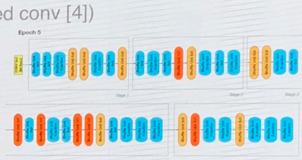
[1] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." ICLR 2015.
[2] Veit, Andreas, Michael J. Wilber, and Serge Belongie. "Residual networks behave like ensembles of relatively shallow networks." NIPS 2016.
[3] De, Soham, and Samuel L. Smith. "Batch normalization biases residual blocks towards the identity function in deep networks." NeurIPS 2020.

Advantages of Large Kernels (cont'd)

MEGVII 旷视

③
- 大卷积核对FCN-based的下游任务提升明显
  - Detection (e.g. deformable conv [1], DetNAS [2])
  - Segmentation (e.g. global conv [3], dilated conv [4])

k	3	5	7	9	11
Score (GCN)	70.1	71.1	72.8	73.4	73.7
Score (Stack)	69.8	71.8	71.3	69.5	67.5

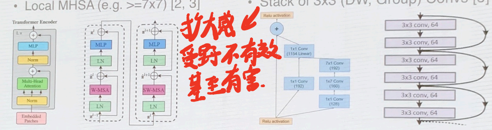


[1] Dai, Jifeng, et al. "Deformable convolutional networks." ICCV 2017.
[2] Chen, Yukang, et al. "Detnas: Backbone search for object detection." NeurIPS 2019.
[3] Peng, Chao, et al. "Large kernel matters--improve semantic segmentation by global convolutional network." CVPR 2017.
[4] Chen, Liang-Chieh, et al. "DeepLab: Semantic image segmentation with deep convolutional nets, atrous…"

Spatial Modeling

MEGVII 旷视

- Vision Transformers
  - Global MHSA [1]
  - Local MHSA (e.g. >=7x7) [2, 3]
- CNNs
  - Large kernels (e.g. >= 5x5) [4]
  - Stack of 3x3 (DW, Group) Convs [5]



扩大感受野不存在
是否有害

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." ICLR 2021.
[2] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." ICCV 2021.
[3] Dong, Xiaoyi, et al. "Cswin transformer: A general vision transformer backbone with cross-shaped windows."
[4] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." AAAI
[5] … "Deep residual learning for … " CVPR

为什么大卷积核CNN没有成为主流？

MEGVII 旷视

- 架构超参 [1, 2]
  - 宽度
  - 深度
  - 输入分辨率
  - 卷积核大小

Large Kernel CNN Design

MEGVII 旷视

数据集太弱？

- 问题：ImageNet的局限性
  - ImageNet分类可能更偏向纹理特征 [1]
  - ImageNet分类任务对感受野要求不高 [2, 3]
- 对策：更强的训练和数据增广策略

[1] Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." ICLR 2019.
[2] Peng, Chao, et al. "Large kernel matters--improve semantic segmentation by global convolutional network." CVPR 2017.

Large Kernel CNN Design (Cont'd)

MEGVII 旷视

- 问题：大卷积核不够高效
- 对策
  - 更浅的结构 [1]
  - 卷积核分解 [2]
  - FFT Conv [3]
  - 稀疏算子 [4]

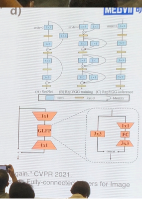
[1] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." ICCV 2021.
[2] Han, Qi, et al. "Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight."
[3] Rao, Yongming, et al. "Global filter networks for image classification."

小结：通向大感受野视觉模型设计

- 采用更强的训练和数据增广策略
- 浅且高效的Spatial算子
- 使用结构重参数化等方法添加架构先验
- 对Spatial和Depth分别设计

Large Kernel CNN Design (Cont'd)

- 问题：大卷积核难以兼顾局部特征
- 对策
  - 结构重参数化方法 [1, 2]



Spotlight

1. Condense V2