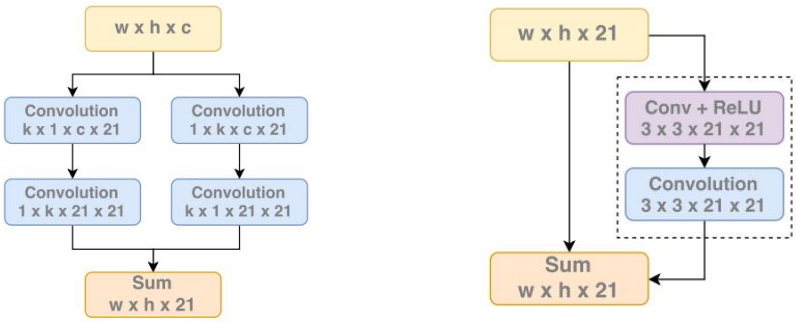
读书笔记

1. Large Kernel Matters - Improve Semantic Segmentation by Global Convolutional Network

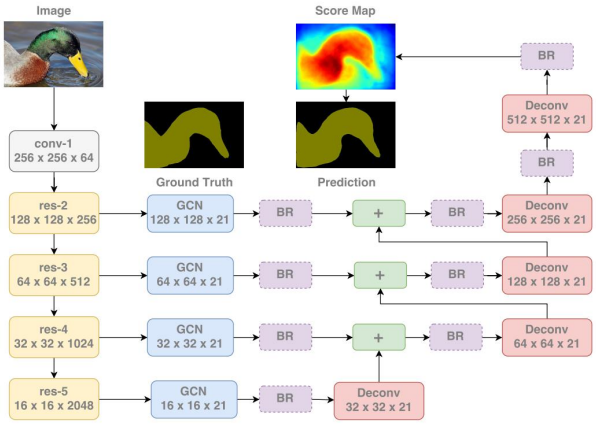
Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, Jian Sun

在作者看来，分类和定位这两个任务是相互冲突的，因为在分类的时候希望特征是移不动的，而定位的时候要求对位置敏感，而在做 segmentation 的时候，正是这两个任务的结合。而对语义切分的主要改进方向一般包括：1. Context Embedding 2.Resolution Enlarging 3.Boundary Alignmen。论文主要强调大的卷积核的重要性，在 PASCAL VOC 2012 数据集上达到 82.2%的 mean-IoU。

该文章认为通过设置大的感受野可以获得特征图与像素之间的紧密关联，并且设置了 Global Convolutional Network 和 Boundary Refinement 两个部件来提高实验的准确率。而所谓的 Global Convolutional Network 就是增大卷积核大小来获得更大的感受野。两个部件的内容如图一所示。整体结构如图二所示。



图一、Global Convolutional Network（左）和 Boundary Refinement（右）



图二、网络整体结构

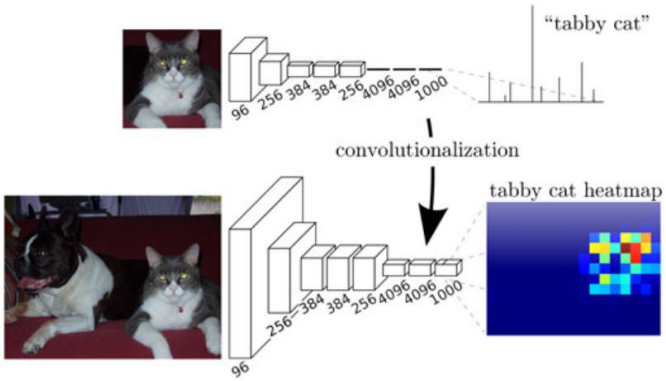
此外，论文里还进行了多种的对比实验，得出的结论为：使用一个大的卷积核可以舒缓分类与定位之间的矛盾，但是模型的准确率增长并不能依靠网络参数的增加，因为过多的参

数不好优化反而会导致效果的不好。

2. Fully Convolutional Networks for Semantic Segmentation

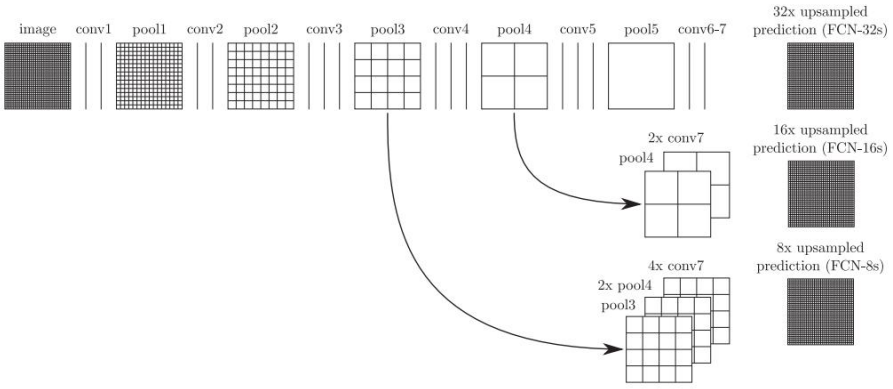
Evan Shelhamer, Jonathan Long, and Trevor Darrell

这篇论文是 CNN 运用到 Segmentation 的鼻祖，对后面的切分任务的影响举足轻重。这篇文章的主要思路是把 CNN 最后的全连接层改为卷积，最后进行上采样。输入一幅图像后直接在输出端直接输出每个像素所属的类别，从而得到一个端到端的方法来实现图像的语义分割。网络结构如图三所示。



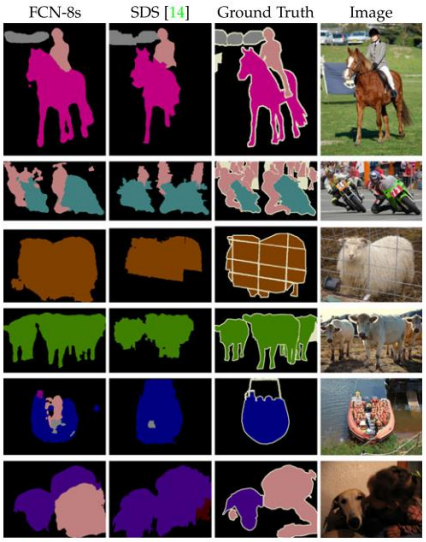
图三、FCN 整体网络结构

另外，由于只使用最后的 1000*1*1 的特征图进行上采样会出现分类分辨率低的情况。为了缓解这个问题，作者还使用了倒数几层的输出和最后的输出做一个融合。结构如图四所示。



图四、FCN fuse 结构

对于 FCN，还是存在着容易丢失较小的目标的问题，如图五中第一行图片中的汽车，和第二行图片中的观众人群。



图五、FCN 存在的问题

3. Rethinking Atrous Convolution for Semantic Image Segmentation

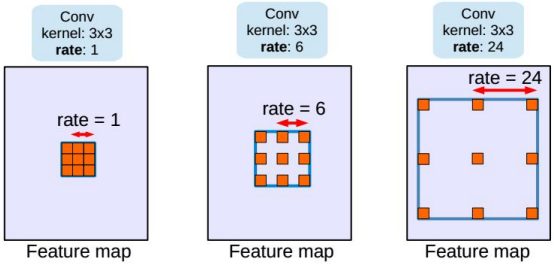
Liang-Chieh Chen, George Papandreou, Florian Schroff, Hartwig Adam

全卷积的深度卷积网络（DCNN）对于语义分割的效果表现的很好。然而，网络中的连续重复使用 max-pooling 和 striding 的组合会显著减少特征图的空间分辨率。对此，这篇论文建议使用带孔卷积（Atrous Convolution）进行代替。

假设一个二维信号，每个位置 i 对应的输出为 y 和卷积核为 w，带孔卷积在输入特征图上的计算如下：

$$\boldsymbol{y}[\boldsymbol{i}] = \sum_{\boldsymbol{k}} \boldsymbol{x}[\boldsymbol{i} + r \cdot \boldsymbol{k}]\boldsymbol{w}[\boldsymbol{k}]$$

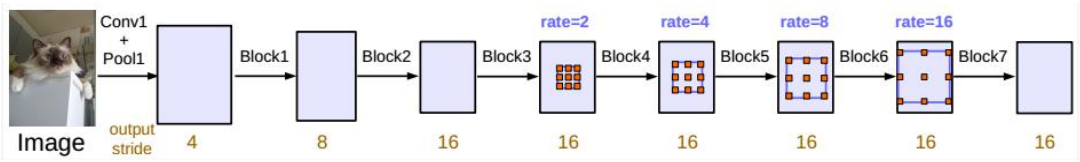
其中孔的比例为 r 对应采样输入信号的步长，这相当于将输入 x 与通过在每个空间维度上两个连续的卷积核值之间插入 r-1 个零点而产生的上采样滤波器进行卷积。标准的卷积是 r=1 的情况，而带孔卷积能够通过改变比例值自适应地修改滤波器的感受域，如图六所示。



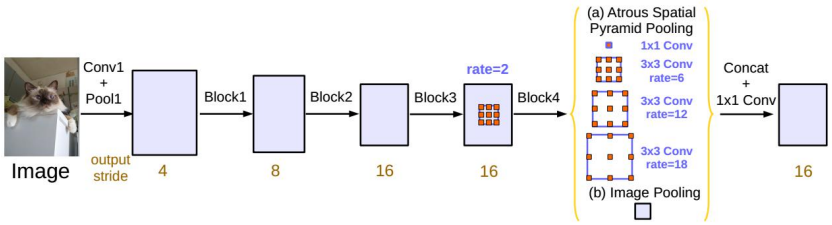
图六、size 为 3*3 的不同比例值的带孔卷积

在这篇论文中，还提出了空间金字塔池化（ASPP）和带有带孔卷积的深度框架。其实，

后者则是只有一条支路的，前者为并行使用多个不同比例值的带孔卷积，以适应不同尺寸的物体。结构如图七所示。



(a)带有带孔卷积的深度框架



(b)ASPP 结构的带孔卷积

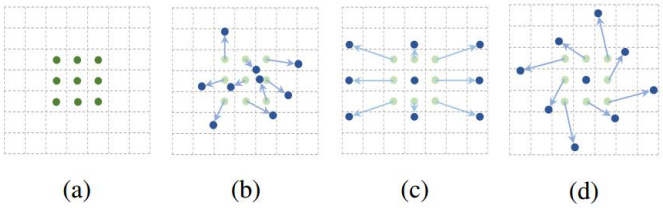
图七、串行与并行的两种带孔卷积结构

4. Deformable Convolutional Networks

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, Yichen Wei

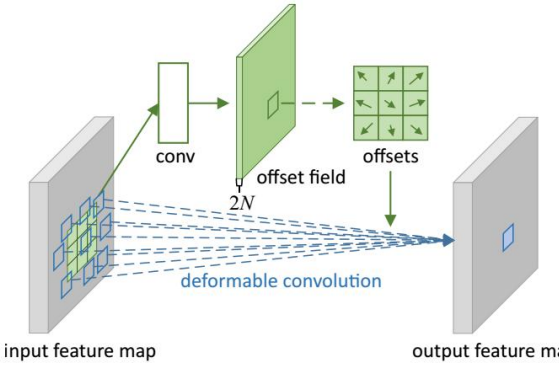
在很多的任务当中，都存在着图像形变的问题，为了缓解这个问题，一般有两个方法：一是数据增强，二是使用一些人工提取的移不变特征，例如 SIFT 等。但这些解决方法是存在其局限的，例如数据增强只能解决预先设定的变化，人工提取的特征难度较高等。

在这篇论文中，则提出了一种可变形的卷积，使得在卷积的时候可以使网络自适应于物体的形变，还可以对图片的关键部分进行更多的关注。可变形的卷积如图七所示。



图七、可变形卷积

其中，图七中为(a)最为普通的卷积核，(c)(d)为卷积核变形的特殊情况，分别为拉伸与旋转。可变形卷积的实现也是非常简单，只需添加一条支路以生成卷积核的偏置量。结构如图八所示。



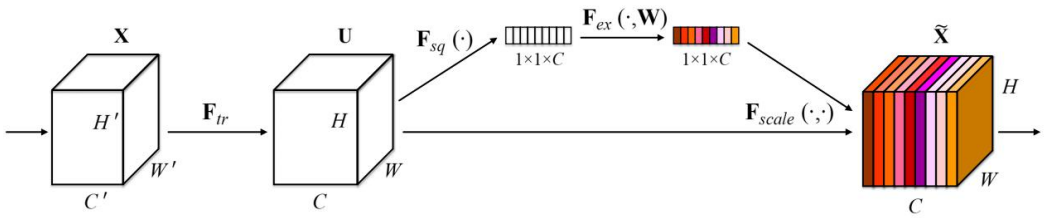
图八、3*3 可变形卷积

可变形卷积可以适用于所有待识别目标具有一定几何形变的任务，而且可以直接由已有网络结构扩充而来，无需重新预训练。虽然只是简单的增加了一个生成偏置量的小部件，仅增加了很少的模型复杂度和计算量，但显著提高了识别精度。例如，在 CityScapes 数据集上，单单是使用可变形卷积神经网络，就将准确率由 70%提高到了 75%。

5. Squeeze-and-Excitation Networks

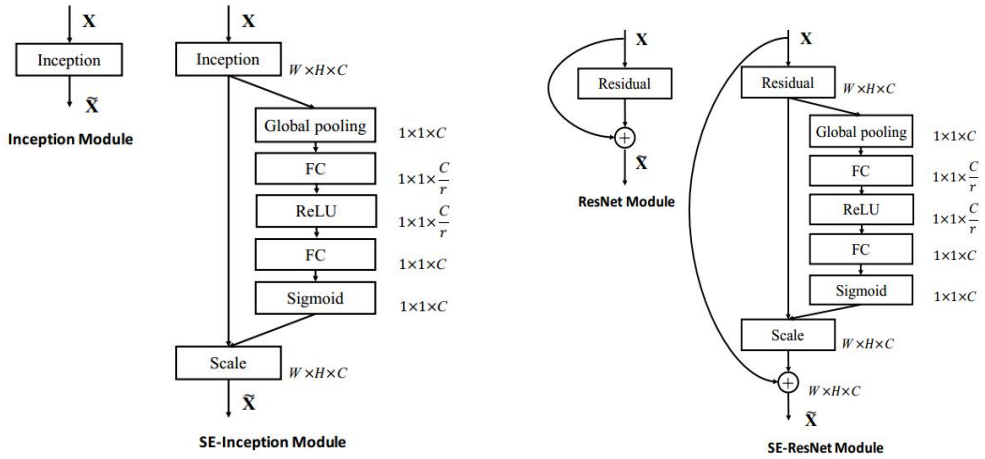
Jie Hu, Li Shen, Gang Sun

在之前的很多工作中，已经在在空间维度上来提升网络的性能，但是他们在所有通道上产生的特征都是等权重的。但是为什么各个通道的权重就是要一致呢？而这篇论文就是从各通道间权重大小入手了。这篇文章显式地建模特征通道之间的相互依赖关系，采用了一种全新的特征重标定策略，通过学习的方式来自动获取到每个特征通道的重要程度，然后依照这个重要程度，去提升有用的特征并抑制对当前任务用处不大的特征。网络的结构如图九所示。



图九、SEnet 网络结构

网络通过增加一个支路，先把特征图的每个通道 squeeze 成一个特征点，然后再把各个特征点拼接成为一条向量在进行全连接，得出各个通道的重要程度（权重），再将其使用在另一条支路上。而且这种想法可以很容易地迁移到已有的网络当中去，如图十所示。另外，该做法在 ImageNet 2017 夺冠。

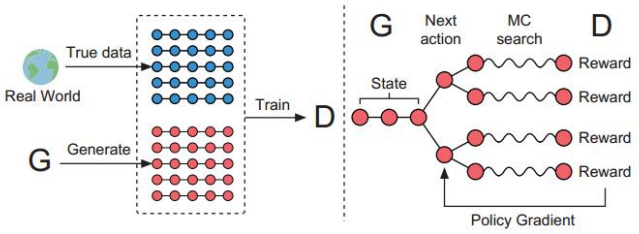


图十、SEnet 的迁移

6.SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient

Lantao Yu, Weinan Zhang, Jun Wang, Yong Yu

在之前的看法中，NLP 仿佛是与 GAN 是互不相容的。因为在语言里面，每一个单词都是映射到一个词向量中，是离散的，不像图片那样是连续的，这就导致了当鉴别器反馈回来信息的时候，生成器无法得到相应的更新，因为，“我们”+0.01 这样的向量是不存在的。为了解决这个问题，这篇论文使用了门特卡罗思想以及 policy gradient 方法，进行 GAN 的训练。网络结构如图十一所示。



图十一、SeqGAN 网络结构

当生成器生成到 t 时，对后面 T-t 个时间点采用蒙特卡洛搜索搜索出 N 条路径，将这 N 条路径分别和已经解码的结果组成 N 条完整输出，然后将 D 网络对应奖励的平均值作为 reward.因为当 t=T 时无法再向后探索路径，所以直接以完整生成结果的奖励作为 reward，而蒙特卡洛采样是指根据当前输出词表的置信度随机采样。

算法如 Algorithm 1 所示。在文中也可以看出生成器生成的语言模型也是比较的鲁棒，而且没有之前使用 MLE 方法的无限循环等的缺陷。

Algorithm 1 Sequence Generative Adversarial Nets

Require: generator policy G_θ; roll-out policy G_β; discriminator D_ϕ; a sequence dataset $\mathcal{S} = \{X_{1:T}\}$

Initialize G_θ, D_ϕ with random weights θ, ϕ.
Pre-train G_θ using MLE on $\mathcal{S}$
$\beta \leftarrow \theta$
Generate negative samples using G_θ for training D_ϕ
Pre-train D_ϕ via minimizing the cross entropy
repeat
  for g-steps **do**
    Generate a sequence $Y_{1:T} = (y_1, \ldots, y_T) \sim G_\theta$
    for t in $1 : T$ **do**
      Compute $Q(a = y_t; s = Y_{1:t-1})$ by Eq. (4)
    end for
    Update generator parameters via policy gradient Eq. (8)
  end for
  for d-steps **do**
    Use current G_θ to generate negative examples and combine with given positive examples $\mathcal{S}$
    Train discriminator D_ϕ for k epochs by Eq. (5)
  end for
  $\beta \leftarrow \theta$
until SeqGAN converges

7.Language Generation with Recurrent Generative Adversarial Networks without Pre-training

Ofir Press, Amir Bar, Ben Bogin, Jonathan Berant, Lior Wolf

这篇论文可以说是在上一篇的论文的基础上，增加了一些新的训练方法，以使结果可以略微提高。这篇论文的主要工作就是使用了 Curriculum Learning(CL), Variable Length (VL)与 Teacher Helping (TH)。

在这篇论文中，使用了两个 RNN 网络分别充当生成器与判别器。简单来说，在这个任务里面，Curriculum Learning 就是逐渐增加判别器与生成器的序列长度，即由 1 到 32 个字符；Variable Length 则是在每个 batch 里面，随机生成的序列长度为 1~32，长度随机；Teacher Helping 则是把生成器的生成的一个字符拼在真实的序列后面，再送到判别器中去。实验结果如图十二所示。（输入为随机噪声）

CL	VL	TH	Samples	%-IN-TEST-n 1	2	3	4
✗	✗	✗	???ccccccccccccccccccccccccccc &&;?x?????++++++?+?+?++++++++++ ?vVVV5--5-?-?-?-?-?-?-?s?-ss{6?	28.8	3.7	0.0	0.0
✗	✓	✗	??nnnnnnnnnnnnnnnnneeeee mfe mf rerrrrrrrr e an e ao e a e ho e "h"p t t t t t t t ' t h e e a	80.6	8.6	0.0	0.0
✓	✗	✗	1x????????????? ?????????????? Bonererennerere ?Sh???????orann unngenngHag g g?e????????????	27.0	7.9	2.0	0.0
✓	✓	✗	The prope prof ot prote was the Wy rronsy ales ale a Claie of th Price was one of the plaids rom	68.1	24.5	4.4	0.5
✗	✓	✓	The increase is a bilday in the Sment used a last give you last She was the intervice is orced t	79.4	44.6	11.5	0.7
✓	✓	✓	Republicans friends like come ti Researchers have played people a The Catalian Office of the docum	**87.7**	**54.1**	**19.2**	**3.8**

图十二、网络实验结果

从实验结过可以看出，单纯的 CL 作用是无效的，而 VL 的加入效果较为明显。而 CL+VL+TH 的效果才是最好的，可以看出这三个训练方法的加入是比较有效的。另外，加入这三个机制之后，生成长句子的质量也有所提高，生成句子如下所示。

Table 3: Samples of length 64 generated by the CL+VL+TH model.

Marks live up in the club comes the handed up moved to a brief d
The man allowed that about health captain played that alleged to
If you have for the past said the police say they goting ight n
However , he 's have constance has been apparents are about home
The deal share is dipled that a comments in Nox said in one of t
Like a sport released not doing the opposition overal price tabl

8.Curriculum Learning

Yoshua Bengio, Jerome Louradour, Ronan Collobert, Jason Weston

这篇文章考虑到人的学习，学的是经过组织的知识，才能学习得更快。那么对应到机器学习，是否能够通过改变学习的顺序或者是说，对知识进行简单的组织，提升机器学习的效果呢？本文经过实验，发现这确实是一个可行的思路，改变学习的顺序，能对学习的速度和质量进行提升。而先易后难是一个有效并且高效的学习策略，本文提出的 curriculum learning 的方法，能够帮助找到更好的局部最小值，以提升结果。

论文还对 Curriculum 进行了定义：

样本 z 的分布概率是 P(z)，权重是 W(z)，那么在某个 step λ（λ范围为[0, 1]），对应的实际值为：

$0 \leq \hat{\lambda} \leq 1$, and $\hat{W}_1(z) = 1$. The corresponding training distribution at step λ is

$$Q_\lambda(z) \propto W_\lambda(z) P(z) \quad \forall z \tag{1}$$

such that $\int Q_\lambda(z)dz = 1$. Then we have

$$Q_1(z) = P(z) \quad \forall z. \tag{2}$$

同时，随着 step 增加（样本数也增加），需要满足：

1.熵递增，以增加训练样本的多样性；

$$H(Q_\lambda) < H(Q_{\lambda+\epsilon}) \quad \forall \epsilon > 0$$

2.权重递增，以提升样本在训练集的作用，从而增加样本数量；

$$W_{\lambda+\epsilon}(z) \geq W_\lambda(z) \quad \forall z, \forall \epsilon > 0.$$

另外，论文还从 supervised learning，形状分类，语言模型等三个实验来说明什么是“easy”的数据，并且从论文中可以看到，curriculum learning 确实是一个可行的方法。

9.Adversarial Learning for Neural Dialogue Generation

Jiwei Li, Will Monroe, Tianlin Shi, Sebastien Jean, Alan Ritter and Dan Jurafsky

这篇论文和之前的工作 SeqGAN 类似，也是采取了增强学习来规避 GAN 在 NLP 中使用的难点，并作出了更多的尝试。

在本任务中的增强学习中，有一个很大的问题，就是我们的估价都是针对一整个回答的，

判别器只会给出一个近似于对或者不对的反馈。这样的的反馈存在一个很大的问题是，即使是很多被判断为假的句子，其中有很大一部分是有效的，如文中的例子“what’s yourname”，人类回答“I am John”，机器回答“I don’t know”。判别器会给出“I don’t know”是有问题的，但无法给出 I 是对的而后面的 don’t know 是错的，而事实上机器没有回答 he/she/you/they 而是 I 本质上是需要一个肯定的正反馈的。判别器只告诉机器对或错，却不告知哪部分对和哪部分错，这对训练带来了很大隐患。所以文中采用了两种方式，第一种是蒙特卡罗算法，而第二种则是使用局部序列来评估。局部序列相当于是蒙特卡罗抽样的一个特殊例子，每次只随机获取一个样本进行更新，解决了蒙特卡罗算法运算量大的问题，但是也存在着梯度指向不准的问题，效果不如蒙特卡罗的好。

另外，在 GAN 的训练过程中，假如识别器比较好，而生成器突然变差了，这就使得网络不断地获取负反馈，导致无法训练。为了减缓这个问题，在训练过程中人工生成的加进去当做是真实的数据，迫使生成器往正常的方向前进，干涉其生成比较真实的数据，这被称为“Teacher Forcing ”。具体算法如图十三所示。

For number of training iterations **do**
. **For** i=1,D-steps **do**
. Sample (X,Y) from real data
. Sample $\hat{Y} \sim G(\cdot|X)$
. Update D using (X, Y) as positive examples and $(X, \hat{Y})$ as negative examples.
. **End**
.
. **For** i=1,G-steps **do**
. Sample (X,Y) from real data
. Sample $\hat{Y} \sim G(\cdot|X)$
. Compute Reward r for $(X, \hat{Y})$ using D.
. Update G on $(X, \hat{Y})$ using reward r
. Teacher-Forcing: Update G on (X, Y)
. **End**
End

图十三、GAN 简要训练算法

实验的结果可以一定程度上体现对抗训练的模型起到了预期的效果。部分实验结果如图十四所示。

10.Wasserstein GAN

Martin Arjovsky, Soumith Chintala, and Leon Bottou

自从 GAN 出现，就存在着训练困难、生成器和判别器的 loss 无法指示训练进程、生成样本缺乏多样性等问题。而这篇文章则是从 GAN 的公式理论推导中找出解决的方法，彻底解决 GAN 训练不稳定的问题，不再需要小心平衡生成器和判别器的训练程度，另外，基本解决了 collapse mode 的问题，确保了生成样本的多样性。而这么大的改善，并不需要改变太多的东西，只要稍微改动以下四点即可：1.判别器最后一层去掉 sigmoid 激活函数 2.生成器和判别器的 loss 不取对数 3.每次更新判别器的参数之后把它们的绝对值截断到不超过一个固定常数 4.不要用基于动量的优化算法，推荐 RMSProp。

具体算法如图十四所示。

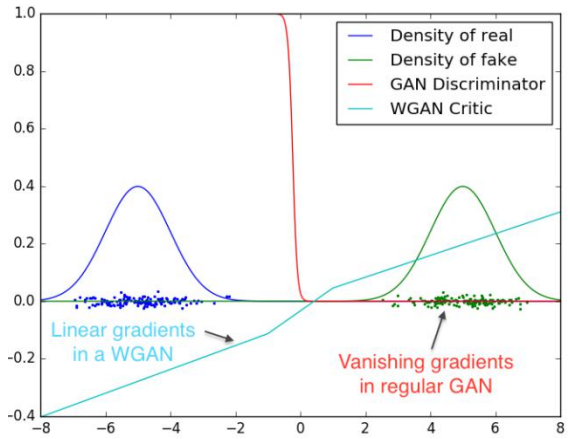
Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

Require: : α, the learning rate. c, the clipping parameter. m, the batch size.
n_{critic}, the number of iterations of the critic per generator iteration.
Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

while θ has not converged **do**
  for $t = 0, ..., n_{\text{critic}}$ **do**
    Sample $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_r$ a batch from the real data.
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
    $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
    $w \leftarrow \text{clip}(w, -c, c)$
  end for
  Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
  $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$
  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end while

图十四、具体算法

该论文引入了 Wasserstein 距离，由于它相对 KL 散度与 JS 散度具有优越的平滑特性，理论上可以解决梯度消失问题。图十五可以看出，用于让一个高斯分布拟合另外一个高斯分布，WGAN 具有更好的特性。其他的一些实验结果也验证了想法。



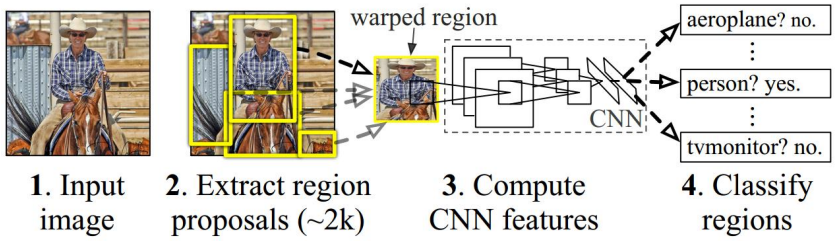
图十五、常规 GAN 与 WGAN 的对比

11.Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik

RCNN 首先将卷积网络应用到检测的任务中去。实现过程大致为，先通过 selective search 算法选择约 2000 个候选区域，这些区域中有我们需要的所对应的物体的 bounding-box，然后对于每一个候选框都 wrap 到固定的大小（227*227），再对每一个处理之后的图片，送进到 CNN 上去进行特征提取，得到每个候选位置的特征图，这些特征用固定长度的特征向量来表示。最后对于每一个类别，我们都会得到很多的特征向量，然后把这些特征向量直接放到 svm 现行分类器去判断，判断当前区域所对应的实物是背景还是所对应的物体类别，每个区域都会给出所对应的得分，再用非极大值抑制进行框的合并，最后就会得到所对应的 bounding-box，再对 bounding-box 进行回归微调，以达到更优的结果。

整个网络结构如图十六所示。



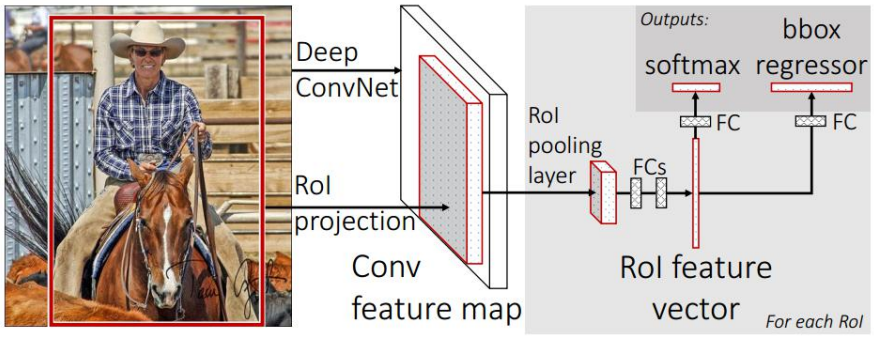
图十六、RCNN 整体结构

但是 RCNN 还是存在着一些问题，例如：1.训练时要经过多个阶段，首先要提取特征微调卷积层，再用线性 SVM 处理候选，计算得到的特征，再进行 bounding-box 的回归 2.训练时间和空间开销大。要从每一张图像上提取大量候选区域图片，还要从每个候选中提取特征，并存到磁盘中。3.测试时间开销大。

12.Fast R-CNN

Ross Girshick

这篇论文主要是对 RCNN 的一些改进。主要贡献在于：1.实现大部分结构的 end-to-end 训练，所有的特征都暂存在显存中，就不需要额外的磁盘空间。 2.联合训练，在 SVM 分类，bbox 回归等过程联合在 CNN 阶段训练。提出了 Multi-task loss，把最后一层的 Softmax 换成两个，一个是对区域的分类 Softmax，另一个是对 bounding box 的回归。3.提出了一个 RoI 层，算是 SPP 的变种，SPP 是 pooling 成多个固定尺度，RoI 只 pooling 到单个固定的尺度。整体网络结构如图十七所示。



图十七、Fast RCNN 的整体网络结构

The RoI pooling layer：对整张图片进行卷积，然后把一开始的那些 RoI 映射到特征图上，然后再对其进行 pooling，生成一个固定大小 ROI 特征向量。而 SPPnet 是使用了多种的尺度来适应输入图片的尺寸差异。

Multi-task loss：

$$L(p,u,t^u,v) = L_{\text{cls}}(p,u) + \lambda[u \geq 1]L_{\text{loc}}(t^u,v),$$

其中，u 是真的 lable，v 是真实的框的坐标，然后，p 是预测出来的类别，t 是预测出来的框的坐标，另外 L_{cls} 则是简单的分类 loss，L_{LOC} 则是

$$L_{\text{loc}}(t^u, v) = \sum_{i \in \{\text{x,y,w,h}\}} \text{smooth}_{L_1}(t_i^u - v_i),$$

in which

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases}$$

作者这样设置的目的是想让 loss 对于离群点更加鲁棒，控制梯度的量级使得训练时不容易跑飞。在讨论中，作者说明了 Multitask loss 是有助于网络的性能提高的。

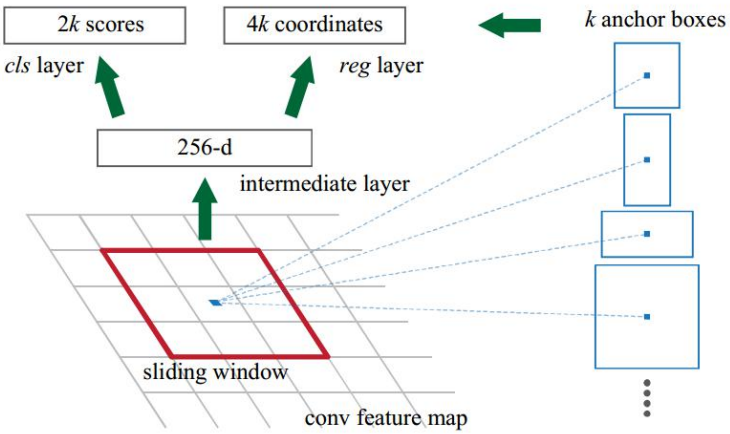
另外，作者还使用了 Truncated SVD 来加速网络的运行。

13.Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun

这篇论文中，创造性地采用卷积网络自行产生建议框，并且和目标检测网络共享卷积网络，使得建议框数目从原有的约 2000 个减少为 300 个，且建议框的质量也有本质的提高。而这篇论文的最大的亮点也在于 RPN 的使用，使前向时间大大减少。

RPN 的核心思想是使用卷积神经网络直接产生候选框，使用的方法本质上就是滑动窗口。RPN 网络结构图如图十八所示，假设给定 600*1000 的输入图像，经过卷积操作得到最后一层的卷积特征图。然后对特征图进行滑窗，滑窗中心点位置，对应预测输入图像 3 种尺度和 3 种长宽比的 anchor。即每个 3*3 区域可以产生 9 个候选框。后边接入到两个全连接层，分别用于分类和边框回归。分类任务包含 2 个元素，用于判别目标和非目标的估计概率。回归任务包含 4 个坐标偏移元素（x,y,w,h），用于确定目标位置。最后根据候选框得分高低，选取前 300 个 region proposal，作为 Fast R-CNN 的输入进行目标检测。



图十八、RPN 网络结构图