论文分享

报告人：郭晓龙

目录

Real-Time, Highly Accurate Robotic Grasp Detection using Fully Convolutional Neural Networks with High-Resolution Images

基于全卷积神经网络与高分辨率图像的实时高精度机器人抓取检测

IEEE ICRA 2019

96.6% detection accuracy

SOTA real-time computation performance on Cornell dataset.

for high-resolution images(6-20ms per 360×360 image)

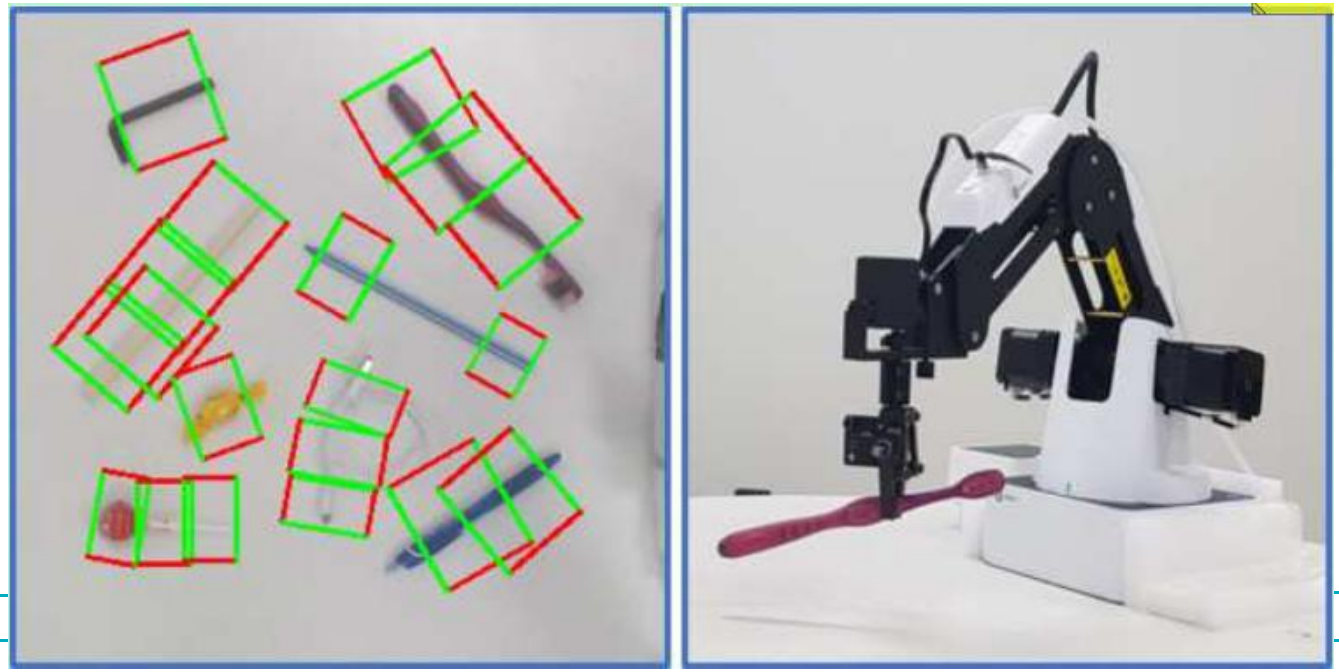
Se Young Chun

Robotic Grasp Detection

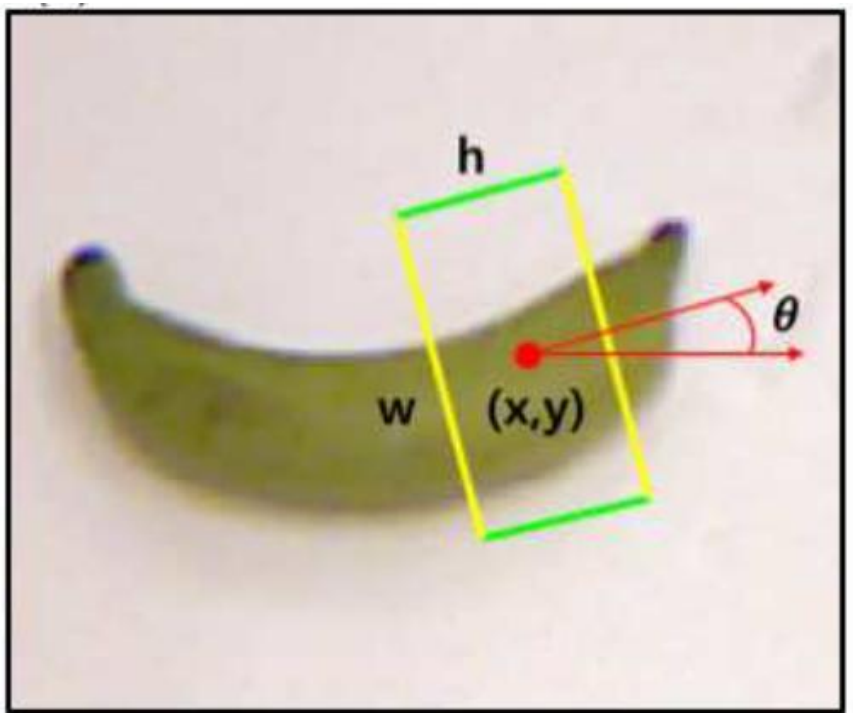
机器人抓取新物体（novel object）已经得到了广泛的研究，但机器人抓取检测（Robotic Grasp Detection）仍旧是个难题。对于机器人来说，如何进行准确的抓取检测、轨迹规划以及可靠地执行，是非常具有挑战性的。

作为机器人抓取的第一步，从成像传感器(如RGB-D相机)中准确、快速地检测机器人抓取，是机器人抓取成功的重要任务。实时机器人抓取检测对于动态环境或动态对象的应用至关重要，因此在保持较高预测精度的同时减少计算时间是十分必要的。

作者提出了一种基于全卷积神经网络(FCNN)的实时机器人抓取检测方法，可以执行多目标、多抓取检测（左图）。并在四轴机器人上，对提出的方法进行了评估，执行对新对象的抓取任务，成功率达到了90%（右图）。



抓取方法的目标是，预测一个机器人抓取的5D表示。即：从一个给定的彩色图像(RGB)或深度图像(RGB-D)，预测一个包含位置(x, y)、方向角θ、钳子打开宽度w、平行夹持板尺寸h的一个5D机器抓取表示：{x, y, θ, w, h}。



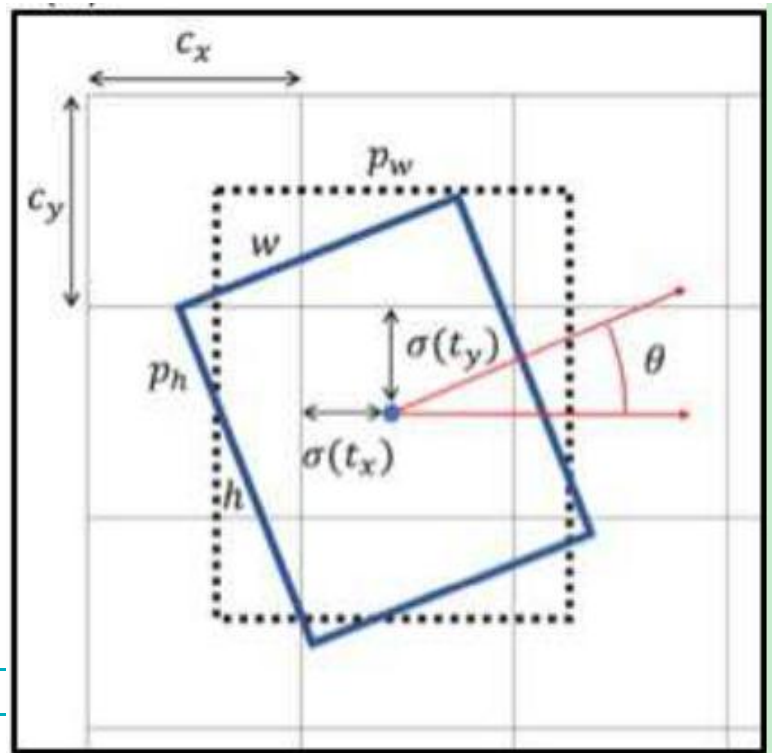
- 角度离散化：θ ∈{0, π/18, . . ., π}.

- 基于摄像机的视觉坐标系下的5D标示，需要转化为实际机器人坐标系下的新的5D抓取标示: {x ˜, y ˜, θ ˜, w˜, h˜},以便机器人执行真实抓取动作。

Problem Description

Parameters

- 每个网格单元的坐标偏移量(x, y)：作者提出的方法不是预测(x, y)图像坐标，而是通过估计每个网格单元(cx, cy)左上角的(x, y)偏移量来预测机器人抓取的位置。
- 每个单元格中的锚框的大小w、h：使用7种大小的默认框，(w, h) ∈ {(0.76, 1.99), (0.76, 3.20),(1.99, 0.76), (1.99, 1.99), (1.99, 3.20),(3.20, 3.20), (3.20, 0.76)}.



Parameters Description

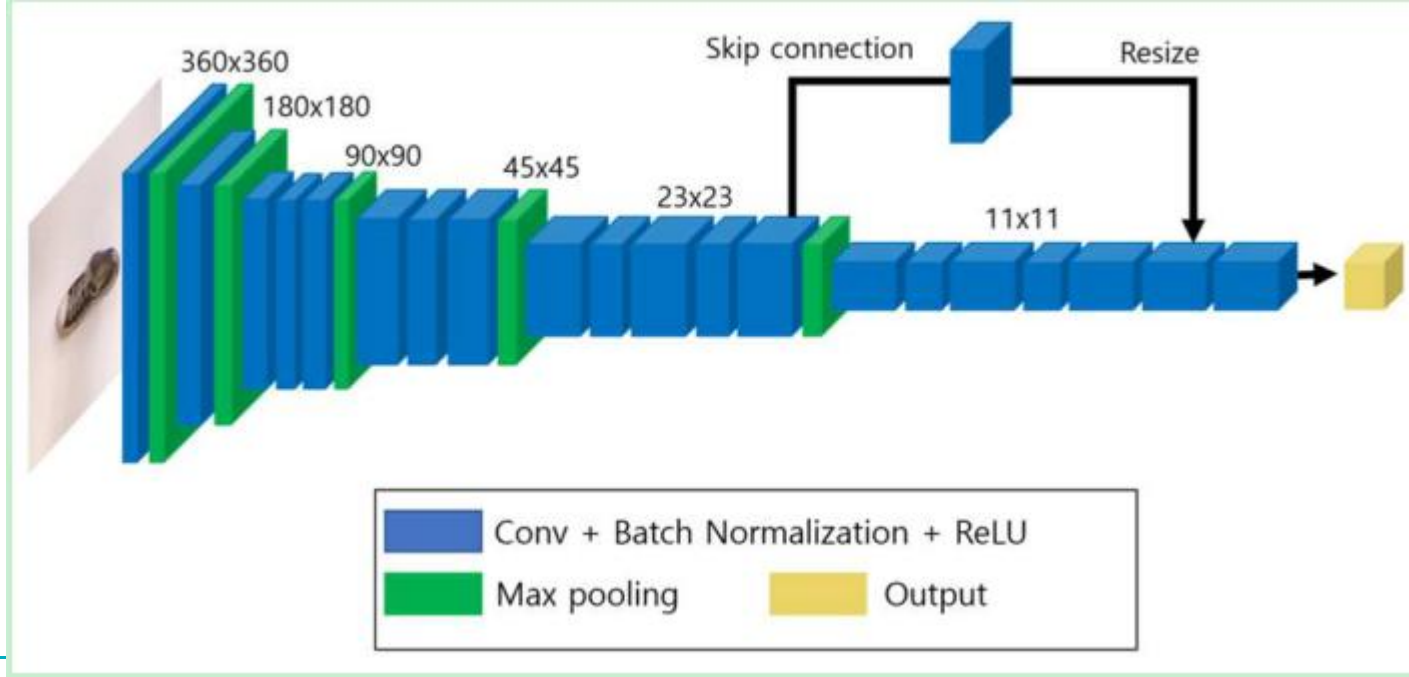
Loss function

作者提出了一种新的机器人抓取检测损失函数：

Grasp probability ground truth = $|P \cap G| / |P \cup G|$;

where ,P is the predicted grasp rectangle, G is the ground truth grasp rectangle, and $|\cdot|$ is the area of the inner set.

网络结构受YOLO9000的启发。



loss function

Loss function

$$
\begin{aligned}
L(t^x, t^y, \theta, t^w, t^h, t^z) = \\
&\lambda_{\text{coord}} \sum_{i=1}^{S^2} \sum_{j=1}^{A} m_{ij}^{obj} [(x_i^g - x_i)^2 + (y_i^g - y_i)^2] + \\
&\lambda_{\text{coord}} \sum_{i=1}^{S^2} \sum_{j=1}^{A} m_{ij}^{obj} [(w_{ij}^g - w_{ij})^2 + (h_{ij}^g - h_{ij})^2] + \\
&\lambda_{\text{prob}} \sum_{i=1}^{S^2} \sum_{j=1}^{A} m_{ij}^{obj} [(z_i^g - z_i)^2] + \\
&\lambda_{\text{class}} \sum_{i=1}^{S^2} \sum_{j=1}^{A} m_{ij}^{obj} \text{CrossEntropy}(\theta_i^g, \theta_i)
\end{aligned}
$$

loss function

Cornell grasp detection dataset

855 images (RGB color and depth) of 240 different objects with the ground truth labels of a few graspable rectangles and a few not-graspable rectangles.

Trainng with a single GPU (NVIDIA GeForce GTX1080Ti), a single CPU (Intel i7-7700K 4.20GHz) and 32GB memory.

Dex-Net 1.0: A Cloud-Based Network of 3D Objects for Robust Grasp Planning Using a Multi-Armed Bandit Model with Correlated Rewards

IEEE ICRA 2016

一个基于云的三维物体网络，使用一个多臂强盗模型和相关的奖励来进行稳健的抓取计划

University of California,Berkeley

简介

基于云的机器人和自动化系统通过网络交换数据和执行计算，而不是在有限的计算和内存中单独运行。使用云（包括大数据）的潜在优势包括:访问更新的图像、地图和产品数据库；访问并行网格计算进行统计分析、机器学习和规划。

作者提出DexNet 1.0（灵巧网络），这是一个新的数据集和相关算法，研究大数据和云计算对稳健抓取规划的规模效应（scale effect）。对于大量可能的对象形状、姿态和环境配置，大规模机器学习抓取可能会表现出与计算机视觉和语音识别类似的规模效应。

为了研究规模效应（scale effect），作者开发了dex-net 1.0，这是一个不断扩增的数据集。目前包括超过10,000个独特的3D对象模型，这些模型被选中来反映库房或家庭中可能遇到的对象，如容器、工具、餐具和玩具等。dx-net还包含大约250万个平行颚式抓取数据，其中的每个对象都标记了最多250个抓取数据，并估计了不确定对象姿态、抓取器姿态和摩擦系数下，每个抓取（数据）的力闭合概率。

作者在谷歌计算引擎上实现了该算法，并将Dex-Net 1.0存储在谷歌云存储上，系统一次可以运行多达1500个实例，进行对象间的相似性度量。减少了3个数量级的运行时间。

Main Contribution

本文的主要贡献在于提出了一种基于具有相关报酬的多臂强盗模型（MAB）算法，该算法通过学习大量的先验知识和三维对象模型来加速鲁棒抓取规划。

抓取规划问题是指，给定一些特定的对象，实现力闭合（force closure）或优化与抓取相关的质量度量（quality metric）。

作者的工作与主动采样抓取的研究紧密相关，从较少的样本中建立抓取质量的统计模型。该算法基于连续相关过程(CCBPs)，这是一个有效的模型，用于预测一个先验数据中每个抓取质量的置信分布。

在这项工作中，作者将MAB模型扩展到三维，研究了在使用Dex-Net 1.0先验数据的情况下，针对高力闭合率的抓取规划的规模效应.

问题定义和描述

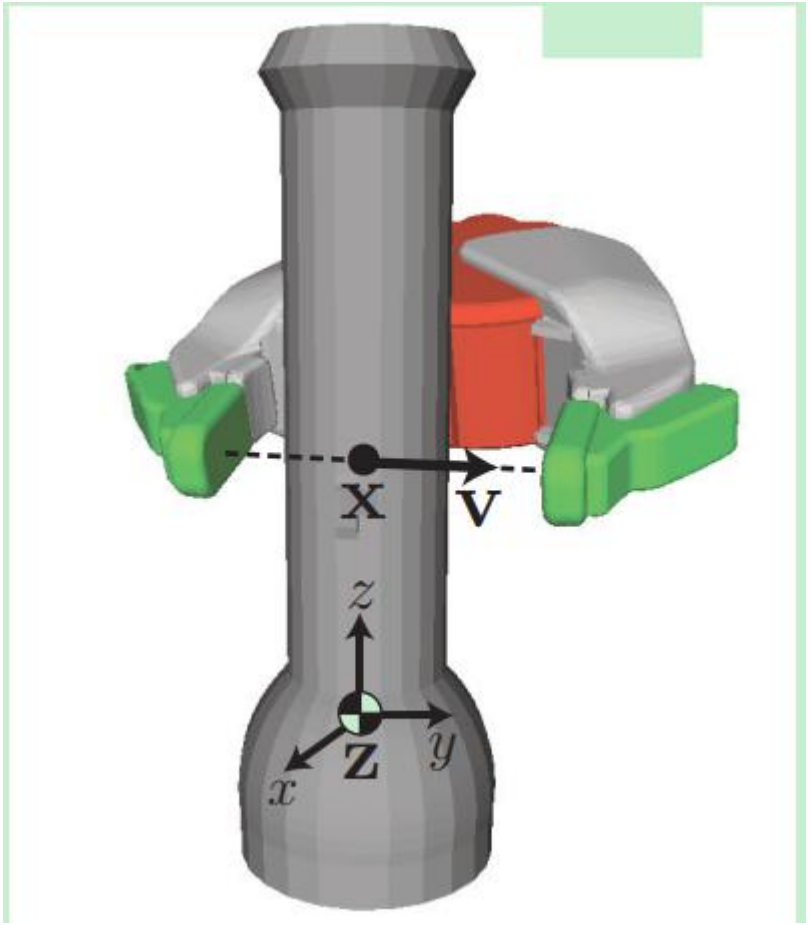
作者考虑了在不确定的目标姿态、夹持器姿态和摩擦系数情况下，在给定三维目标模型和平行爪夹持器下的稳健抓取规划问题。

假设标称物体的形状是一个符号距离函数 (SDF) $f : R^3 \rightarrow R$，它在物体表面为零，在物体外部为正，在物体内部为负。

- 抓取和对象参数化
- 不确定性的来源
- 接触模型
- 质量指标(Quality Metric)
- 整体优化目标(Objective)

DEFINITIONS AND PROBLEM STATEMENT

A.抓取和对象参数化



g = (x， v)——平行颚式抓取
x ∈ R^3 ——颚式抓取器的重心
v ∈ S^2——抓取器趋近方向或轴

S = {y ∈ R^3 |f(y) = 0}——物体的表面

G = { g = (x， v)| x ∈ R^3, v ∈ S^2} ——表示所有抓点的空间

z——物体质心
A——所有SDFs的空间用于封闭和紧凑的表面，指定所有关于，以物体质心z为中心，并沿着S的主轴方向为参考系的点。

H = { O = {z， f(·)} | z ∈ R3, f ∈ A} ——表示所有物体的空间

M = G × H ——所有平行颚抓取和物体的模空间。

Grasp and Object Parameterization

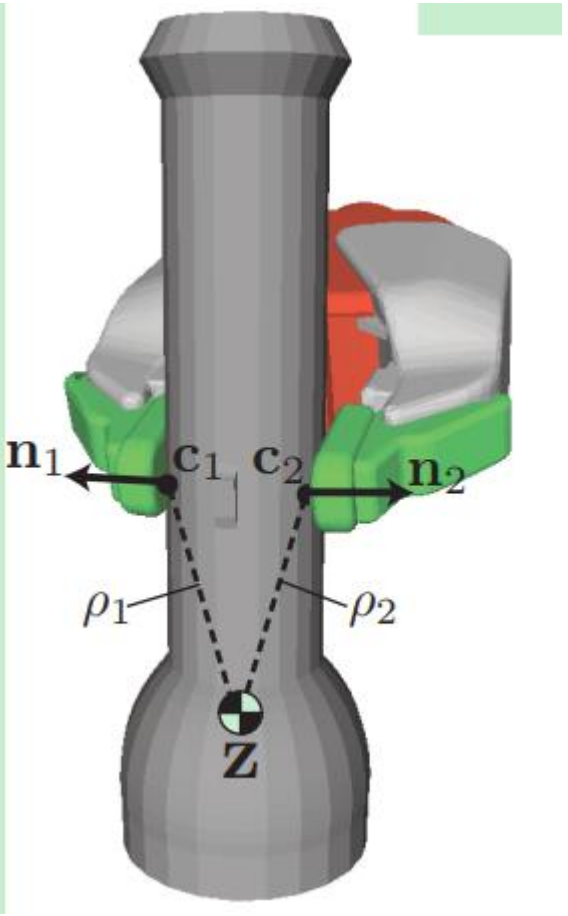
B.不确定性的来源

作者假设在目标位姿、夹持器位姿和摩擦系数上存在高斯分布，以分别对模型误差进行配准、机器人标定或材料属性分类。

ξ——物体位姿 （李代数分布）
ν ——抓取器位姿（高斯分布）
γ——摩擦系数（高斯分布）

以上三者都是从各自分布中抽取的随机变量样本。

C.接触模型



在一个物体O上，给定一个抓取g，ξ、ν、γ分别为物体位姿、抓取器位姿、摩擦系数，然后计算针对抓取g的摩擦锥和转矩——W。

C_1、C_2为触点
n_1、n_2是法向量
Z为物体质心

Contact Model

D.质量指标

在这项工作中，作者使用力闭合的概率(PF)，作为抓取质量度量。

$F \in \{0, 1\}$——力闭合发生的概率

一个在物体O上的抓取g,一个力闭合事件的概率——$PF(g, O) = P(F = 1 \mid g, O, \xi, \nu, \gamma)$.

计算力闭合概率PF：

一个抓取$g \in G$,一个物体$O \in H$,以及随机物体位姿、抓取器位姿、摩擦系数——ξ，ν，γ。首先计算可能的接触扳手的集合——W，Then $F = 1$ if $0 \in \text{Conv}(W)$, where $\text{Conv}(\cdot)$ denotes the convex hull.

E.优化目标

在一个合算的、尽可能大的样本采样空间T内，找到一个抓取g*，使得力闭合率PF最大化。

为了尽可能快地执行这个操作，作者对所有采样点的PF求和求最大值。鉴于在连续空间最大化G的计算开销过于昂贵,过去的工作通过评估一组离散的K候选抓取Γ= $\{g_1,\ldots\ldots, g_k\}$与蒙特卡罗集成或多臂强盗模型(MAB)，解决了这一问题。作者拓展了MAB模型，利用grasps和Dex-Net1.0中3D对象之间的相似性先验信息，来减少搜索空间中样本的数量。

作者还将多视点卷积神经网络(MV-CNNs)，一种最先进的三维形状分类方法，用于有效地检索相似的三维物体。

实验表明，使用先验数据（Dex-net 1.0）可以显著提高稳健抓取规划的质量和复杂程度。

Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metric
基于综合点云和分析抓取度量的深度学习鲁棒抓取规划

University of California,Berkeley

Robotics: Science and Systems 2017

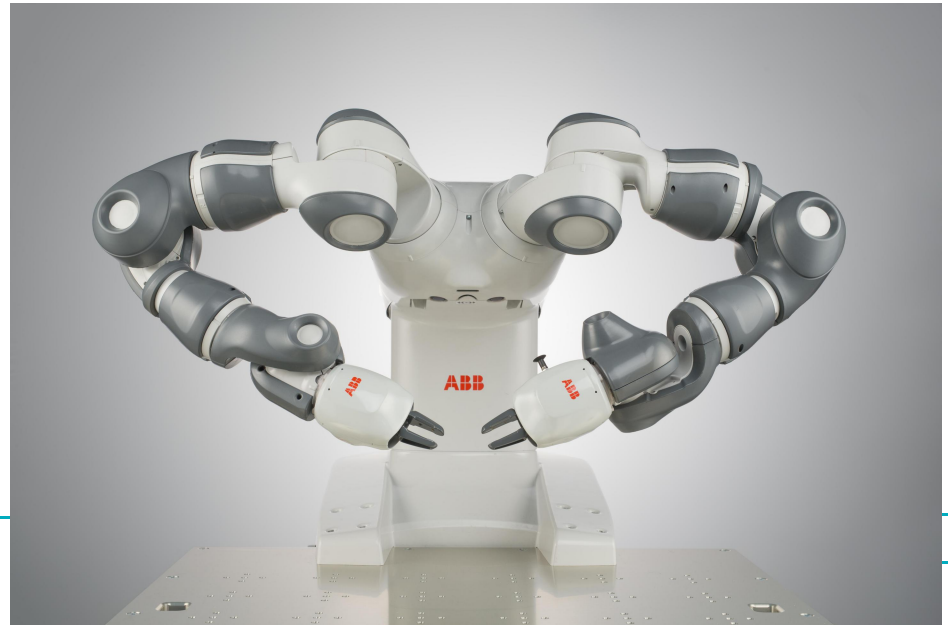
一个抓取质量卷积神经网络(GQ-CNN)架构，该架构可以从点云预测抓取的鲁棒性，并将其训练在Dex-net 2.0上，该数据集包含670万个点云、平行颌抓取和健壮的抓取指标。

简介

为了减少机器人抓取规划所需深度学习数据的收集时间，作者从Dex-Net 1.0的（桌面环境）数千个随机摆放的3D模型，包括670万个点云（point clouds）、抓取（grasps）和分析抓取质量指标（analytic grasp metrics）的合成数据集（Dex-Net 2.0）中进行探索训练。

作者使用得到的数据集Dex-Net 2.0来训练一个抓取质量卷积神经网络(GQ-CNN)模型，该模型可以快速预测深度图像中抓取成功的概率，其中抓取指定为夹持器相对于RGB-D传感器的平面位置、角度和深度。

作者在ABB YuMi上进行了超过1000次的实验。仅使用Dex-Net 2.0的数据训练的GQ-CNN可用于规划0.8秒内的抓取，对8个已知的具有对抗性几何形状的对象的成功抓取率为93%。在包含10个新型刚性对象的数据集上也有最高的成功率，在包含40个新型家居对象的数据集上，其精确度达到99%(在69个被归类为稳健的抓取中，只有一个假阳性)。



Contribution

在本文中，作者在一个包含平行颚抓取(parallel-jaw grasps)、抓取质量指标(grasp metrics)、点云(point clouds)的庞大合成数据集上训练一个深度卷积神经网络(CNN)，利用稳健抓取和图像形成的解析模型，直接从深度图像(depth images)预测抓取成功与否。

- 生成一个关联670万个点云和分析抓取质量度量数据集Dex-Net 2.0；
- 利用期望的质量(expected epsilon quality)作为监督，训练了一个抓取质量卷积神经网络(GQ-CNN)模型,对深度图像中的鲁棒抓取进行分类。其中，每一个抓取（grasp）都被指定为一个相对于相机的平面姿态和深度；
- 一种抓取规划方法，对反方向抓取(antipodal grasp)候选对象进行采样，并用GQ-CNN对其进行排序；

使用ABB YuMi机器人抓取桌面上单个物体，在进行1000多次物理试验后，作者发现，Dex-Net 2.0 抓取规划方法，比基于注册的方法快3倍。抓取训练时已知的物体，达到93%的成功抓取率。在抓取新物体时，也有很好的效果。在40个室内物体的抓取集合上，抓取的准确率达到99%。

主要贡献

给定一个抓取对象和环境可达性约束，抓取规划寻找一个最大限度地获得成功(或质量度量）的夹具配置。

鲁棒性抓取规划(RGP)方法最大限度地提高了抓取的鲁棒性，即在感知和控制不确定性下分析度量（analytic metric）的期望值。

这通常涉及在数据库的3D对象模型上标记抓取（grasp），并使用健壮度量指标（ robust metrics），如力闭合概率（probability of force closure）或位姿误差健壮度量指标（the pose error robust metric ）。

作者对Dex-Net 1.0做了一个主要的扩展，将合成点云与鲁棒性抓取联系起来，并训练一个卷积神经网络来映射点云和候选抓取，用于估计抓取鲁棒性。

问题描述

使用深度相机获取放置在桌面上的模拟刚体的点云，基于点云设计一个鲁棒的平面平行爪抓取（a robust planar parallel-jaw grasp ）。作者学习了一个函数，该函数以候选抓取和深度图像为输入，输出鲁棒性估计，或在传感和控制的不确定性下成功的概率。

A.问题假设

假设一个几何形状已知的平行板夹持器、一个已知内建特性的深度相机、一个在平面上的模拟的刚性物体、以及用深度相机拍摄的单视图(2.5D)点云。

B.数学定义

➢状态

$x = (O, T_o, T_c, \gamma)$——描述摄像机和物体的状态

O——物体的几何性质和质量性质

T_o，T_c——物体和相机的3D姿态

$\gamma \in R$——物体和夹具之间的摩擦系数

➢抓取

$u = (p, \phi) \in R^3 \times S^1$——在三维空间的一个平行颚抓取(a parallel-jaw grasp)

$p = (x, y, z) \in R^3$ ——中心（center）

$\phi \in S^1$ ——平面上的一个角

问题描述

点云

$y = R_+^{H \times W}$——一个H * W 的深度图标示的2.5D点云，这种点云是使用一种已知内部物理学的相机获取的。

➢鲁棒分析抓取度量

S(u， x) ∈ {0, 1} ——是一个二进制值的抓取成功度量，例如力闭合率。

p(S， u， x， y) ——在感知和控制方面，抓取成功、抓取、状态和点云建模不精确，是一个联合分布。

ps : 一个抓取的鲁棒性可以视为：给定观测值下，度量的期望；或者在传感和控制的不确定性下成功的概率。

Q(u， y) = E [S|u,y]

C.目标

目标是，在许多可能的抓取、物体、图像的基础上，学习一个鲁棒性函数$Q_{\theta^*}(u， y) \in [0, 1]$。

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} E_{p(S,u,x,y)}[L(S, Q_\theta(u, y))]$$

L ——cross-entropy loss function

Θ——定义了GQ-CNN网络的参数

对于所有可能的抓取和图像来说，存在$\theta \in \Theta$,使得$Q_\theta = Q$

问题描述

所估计的鲁棒函数，可以用在一个候选抓取集合上的抓取策略中，从而最大化（maximizes）Q_{θ^*}。这个个候选抓取的集合为：

$\pi_\theta(y) = \text{argmax}_{u \in C} Q_\theta(u, y)$，

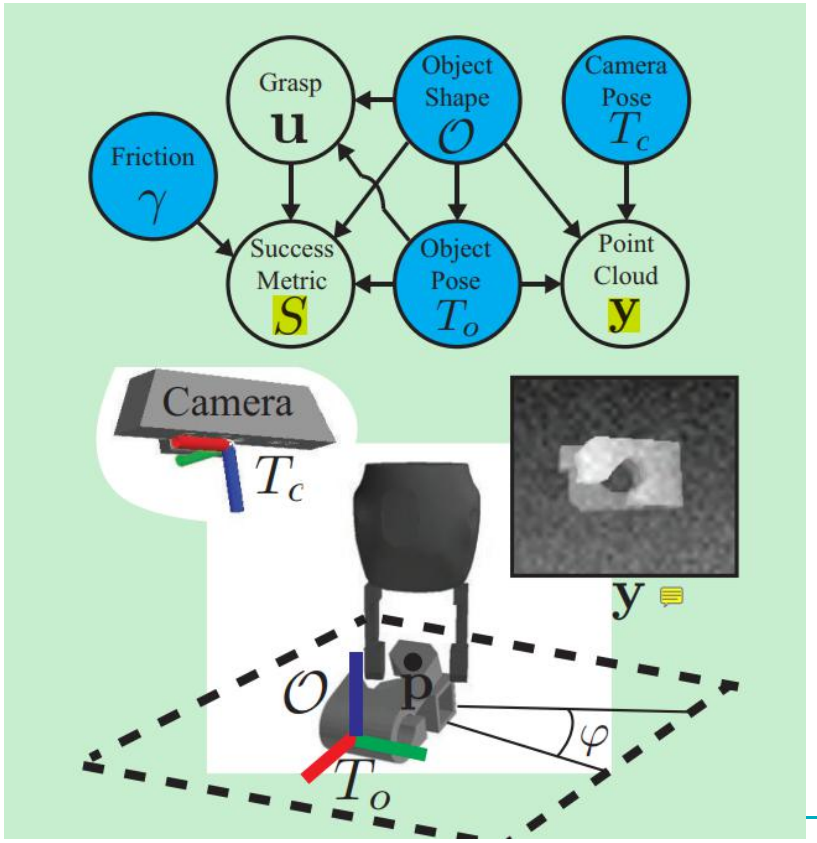
C指定了对可用的grasps集的约束，如碰撞或运动学可行性。

学习Q而不是直接学习策略允许我们强制执行特定于任务的约束，而不必更新所学习的模型。

抓取鲁棒性函数的求解

作者认为，抓取鲁棒性函数的求解具有一定的挑战性。

首先，可能需要大量的样本来近似大量可能对象的期望。作者通过生成dex - net 2.0来解决这个问题，该2.0是一个包含670万个合成点云、平行颌抓取以及从下图中的图形模型中采样的1500个3D模型的稳健分析抓取指标的训练数据集。



其次，在大型对象数据集上的点云、grasps和度量之间的关系可能很复杂，很难用线性或内核化模型来学习。因此，作者开发了一个抓取质量卷积神经网络(GQ-CNN)模型，该模型对深度图像中的稳健抓取姿态进行分类，并在dex-Net2.0的数据上对模型进行训练。

蓝色节点——状态表示中包含的变量。
O——均匀地分布在一组离散的对象模型上。
To——物体姿态，分布在物体的稳定姿态和一个平面的有界区域上。
Grasps u = (p， ϕ)——采用podality constraints从物体表面均匀采样。
y——基于摄像机位姿Tc、目标形状和位姿的三维网格所生成一个合成的2.5D点云，并利用乘法和高斯过程噪声对其进行扰动。

目标求解

基于点云的桌面物体的鲁棒平行爪抓取图模型。

抓取鲁棒性函数的求解

A.图模型

作者使用目标函数的样本近似来估计Q_{θ^*}，使用独立同分布样本$(S_1, u_1, x_1, y_1), \ldots, (S_N, u_N, x_N, y_N) \sim p(S, u, x, y)$,该分布来自于针对图像、抓取、成功度量所生成的图模型。

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}(S_i, Q_\theta(\mathbf{u}_i, \mathbf{y}_i)).$$

p(S, u, x, y)——来自于4个分布

状态分布—— p(x) = p(γ)p(O)p(To|O)p(Tc)
一个观测模型——p(y|x)
一个抓取候选模型——p(u|x)
抓取成功的分析模型——p(S|u, x)

抓取鲁棒性函数的求解

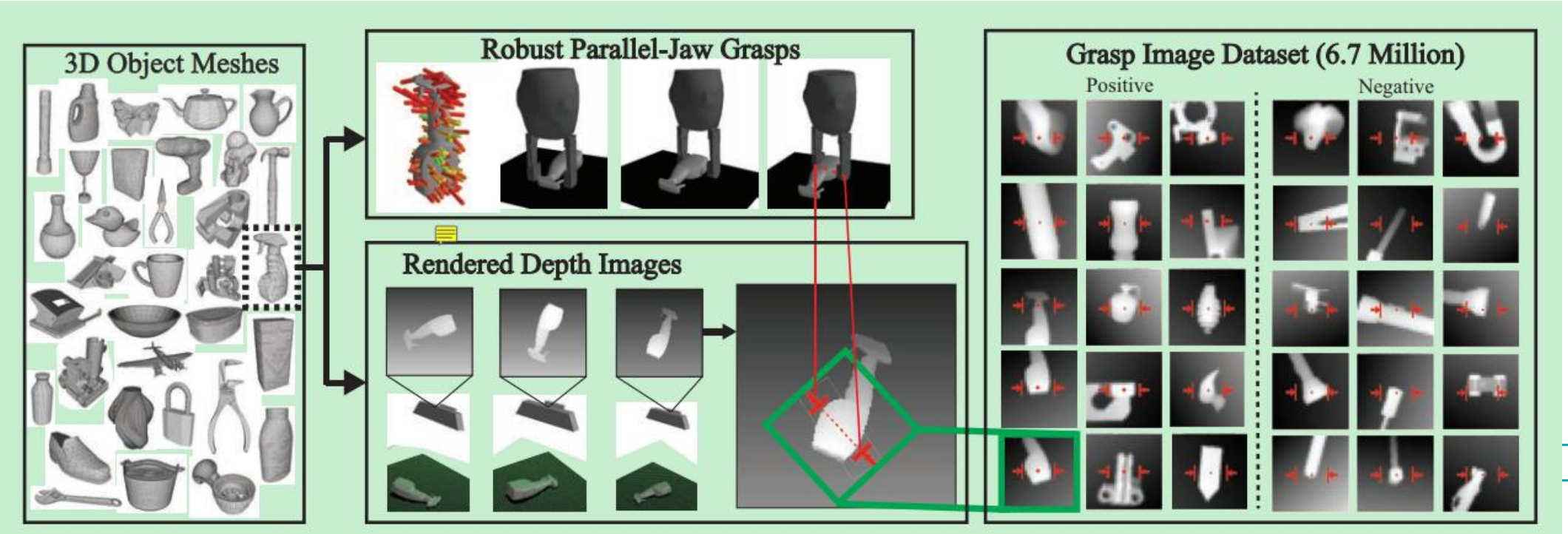
2）数据集

Dex-Net 2.0包含生成的670万个数据点。包括：

●3D模型

●Parallel-Jaw抓取——每个物体都有一组多达100个平行颚式抓取，对于每个抓取，作者使用蒙特卡罗采样来评估在目标位姿、抓取位姿和摩擦系数不确定性下的期望质量EQ。

●点云数据——每个对象还配有一组2.5D的点云(深度图像)，用于每个对象的稳定位姿，相机位姿和平面对象位姿根据之前描述的图形模型进行采样。



用于生成训练数据集的Dex - Net 2.0的pipline。

抓取鲁棒性函数的求解

完整的数据集包含超过670万张抓取图像。

●对于每个对象，作者采样了数百个平行颌抓取来覆盖表面，并使用采样来评估稳健分析抓取指标。

●对于每个稳定的姿态的对象，作者联想到一组抓取是正交于桌子和无碰撞的，对于每个给定的夹具模型。

●对平面对象位姿和相机位姿进行均匀随机采样，在每个稳定位姿中渲染每个对象的点云。

●对于给定的稳定姿态的每个抓取都与渲染图像中的像素位置和方向相关。

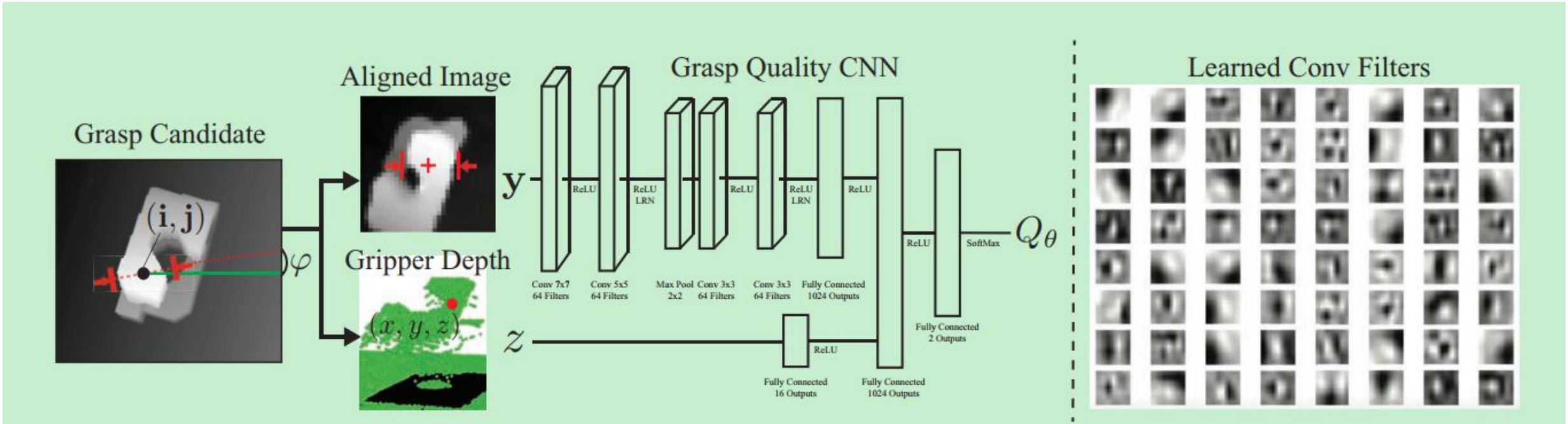
●将每张图像进行旋转、平移、裁剪和缩放，使抓取像素位置与图像中心对齐，使抓取轴与图像中间行对齐，从而得到32×32的抓取图像。

B.抓取质量神经网络

1)架构

GQ-CNN以相对于相机z的抓取器的深度和以抓取中心像素v = (i,j)为中心，与抓取轴方向ϕ对齐的深度图像作为输入，即把深度图像和夹持深度输入GQ-CNN中，估计抓取的鲁棒性。GQ-CNN大约有1800万个参数。定义了参数集 Θ，用于表示抓取鲁棒性函数Q_θ

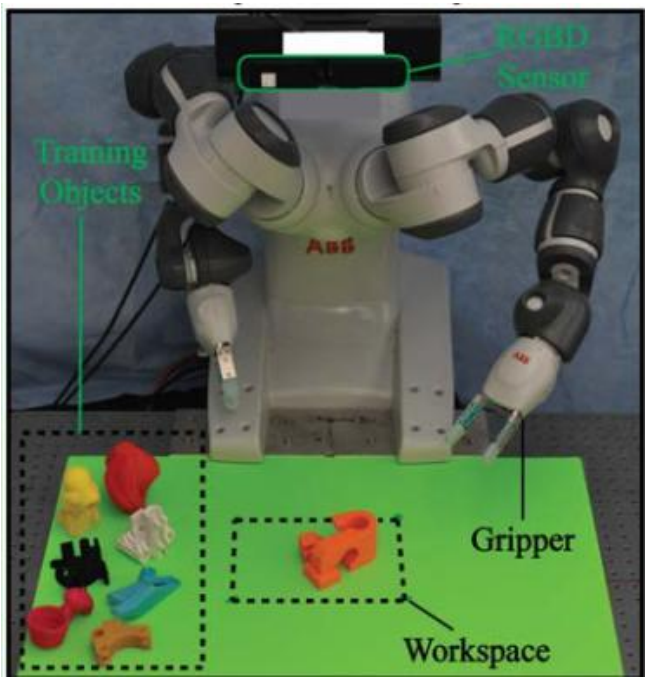
GQ-CNN



2)训练集

GQ-CNN训练数据集是通过将相对于渲染深度图像的像素v、方向ϕ和深度z 与抓取相关联 而生成的。作者利用相机位姿T_c将抓取转换成相机的参考系，并将三维抓取的位置和方向投影到相机的成像平面上，从而计算出这些参数。然后就可以转换所有（图像，抓取配置）对儿，为一个以v为中心、以ϕ为方向的单一的图像。

3)优化
作者利用随机梯度下降和动量的反向传播对GQCNN的参数进行了优化。
作者使用批处理大小为128的TensorFlow实现的模型，动量项0.9，步长0.95的指数衰减进行训练。
在NVIDIA GeForce 1080上训练大约需要48小时。



请 您 指 导

谢谢~