堆叠式Capsule自编码器

摘要

一个物体可以看作是一组相互关联的几何组成部分，因此，一个明确利用这些几何关系来识别物体的系统，在应对视点的变化上理应具有健壮容错性，因为内在几何关系的视点是不会有变化的。我们描述了一个无监督版本的胶囊网络，这里有一个观察所有部分的神经元编码器（neural encoder），它被用来推断目标胶囊的存在和姿态。该编码器通过解码器反向传播来训练，使用混合式姿态预测的方案来预测每个已经发现的部件的姿态。同样是使用神经元编码器，通过推断部件及其仿射变换，可以直接从图像中发现具体的部件。换句话说，每个相应的解码器图像像素建模，都是仿射变换部分做出的混合预测结果。我们通过未标记的数据习得物体及其部分胶囊，然后再对物体胶囊的存在向量进行聚类。最后，该成果在 SVHN 上获得了最先进的无监督分类结果 （55%），同时在 MNIST 上获得了接近最先进的分类结果（98.5%）。

1导言

卷积神经网络比没有权重共享的网络工作得更好，因为它们具有归纳偏置项：如果局部特征在一个图像位置有用，那么同样的特征在其他位置也可能有用。通过多比例，多尺度，仿射自由度等方式进行视角变换重复特征，是很有诱惑力的，但这也会很快地产生繁琐的高维特征映射。

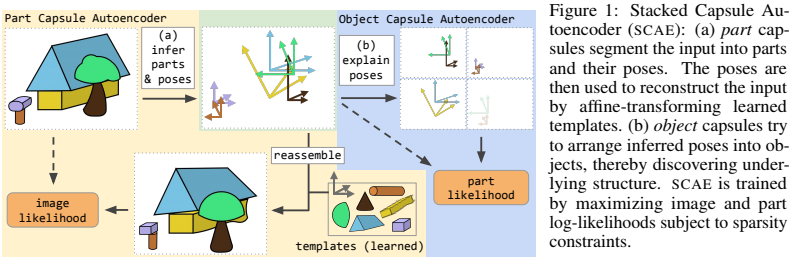
在非平移自由度上，重复特征的另一种方法是明确得到整个对象的自然坐标和每个部分的自然坐标之间的转换方案。计算机图形学依赖于这样的“整体→部分”的坐标变换，以视角不变的方式表示整体对象的几何形状。强有力的证据还表明，人类的视觉与传统的卷积神经网络时不同的。它也依赖于坐标系统：将一个不熟悉的坐标框架强加在一个熟悉的物体上，使其难以识别物体或其几何形状(Rock，1973；Honton，1979)。

神经系统可以在整体，部分，视角的转换之间的学会推理，但是每个转换可能需要不同的表示。对象-部分-关系(OP)是视角不变的，并且由学习的权重自然编码。对象或部分与观察者的关系随着视点而变化(它是视点不变的)，并且自然地使用神经激活来编码。通过这种表示，单个对象的姿态由其与观察者的关系来表示。因此，与中枢神经系统不同，代表一个单一的物体并不需要在整个空间复制神经激活。它只处理两个(或更多)不同的实例

这项工作是在谷歌大脑实习期间完成的。

这可以解释为什么当物体不可见时，要获取关于物体的感知知识，需要用特定的视点来创建物体的心理图像。

预印本。正在审查中。



需要模型参数和神经激活的空间副本。

在本文中，我们提出了堆叠胶囊自动编码器(SCAE)，它有两个阶段(图1)。第一阶段，零件胶囊自动编码器(PCAE)，将图像分割成组成部分，推断它们的姿态，并将每个图像像素重建为转换后的零件模板的像素的混合物。第二个阶段，对象胶囊自动编码器(OCAE)，试图将发现的零件及其姿态组织成一组较小的对象，这些对象可以使用对每个零件单独的预测混合来解释零件姿态。每个物体胶囊通过将其姿态(物体-观察者-关系(OV))乘以相关的物体-部分-关系(OP)来为这些混合物贡献成分。

堆叠胶囊自动编码器(第2部分)在未标记数据上训练时，捕捉整个对象及其部分之间的空间关系。对象胶囊的存在概率向量趋向于形成紧密的聚类，当我们为每个聚类分配一个类时，我们在SVHN (55%)和MNIST (98.5%)上获得了无监督分类的最新结果，通过学习少于300个参数，这可以分别进一步提高到67%和99%。我们还在CIFAR10上展示了有希望的概念验证结果(第3节)。我们在第4节中描述相关工作，并在第5节中讨论我们工作的含义和未来方向。

2个堆叠胶囊自动编码器(SCAE)

将图像分割成多个部分并不容易，所以我们从提取像素和部分发现阶段开始，开发星座胶囊自动编码器(CCAE)(第2.1节)。它使用二维点作为部分，它们的坐标作为系统的输入。CCAE学习将点集建模为熟悉星座的排列，每个星座都通过独立的相似变换进行了变换。CCAE学会了将单个点分配给它们各自的星座——而事先不知道星座的数量或它们各自的形状。接下来，在第2.2节中，我们开发了零件胶囊自动编码器(PCAE)，它学习从图像中推断零件及其姿态。最后，我们将与CCAE非常相似的对象胶囊自动编码器(OCAE)堆叠在PCAE的顶部，形成堆叠胶囊自动编码器(SCAE)。

2.1星座自动编码器(CCAE)

让{xm | m = 1，.。。，M}是一组二维输入点，其中每个点都属于如图2所示的星座。我们首先用集合转换器(Lee等人，2019)——一种基于注意机制的排列不变编码器h帽——将所有输入点(扮演部分胶囊的角色)编码成K个对象胶囊。对象胶囊k由胶囊特征向量ck、其存在概率ak ∈ [0，1]和3 × 3对象-观察者关系(OV)矩阵组成，该矩阵表示对象(星座)和观察者之间的仿射变换。请注意，每个对象胶囊一次只能代表一个对象。每个目标胶囊使用单独的多层感知器(MLP) h部分k从胶囊特征向量ck预测N ≤ M部分候选。每个候选项包括给定候选部分存在的条件概率ak，n ∈ [0，1]，相关的标量标准偏差λk，n，以及3 × 3对象-部分关系矩阵，该矩阵

部件胶囊的类型可以确定物体的哪个部件(如果有的话)有助于用于对已经发现的部件的姿态建模的混合物

2

图2:在不同的位置、比例和方向，对属于最多三个正方形和三角形星座的点进行无监督分割。该模型被训练来重建CCAE混合模型下的点(顶行)。底部一行根据混合模型中具有最高后验概率的父代给点着色。最右边的列显示了一个失败案例。请注意，模型使用点集而不是像素作为输入；我们使用图像只是为了形象化星座排列。

表示对象胶囊和候选零件之间的仿射变换。候选预测k，n由目标胶囊OV和候选运算矩阵的乘积给出。然后，我们将每个输入部分建模为高斯混合，其中k，n和λk，n是各向同性分量的中心和标准偏差。参见图1和图5进行说明；正式描述如下:

OV1:K，c1:K，a1:K = hcap(x1:M)编码对象胶囊参数，(1) OPk，1:N，ak，1:N，λk，1:N = hpart k (ck)解码ck的候选参数，(2) Vk，n = OVkOPk，N解码一个部分姿态候选，(3) p(xm | k，n) = N (xm | k，N，λk，N)将候选转换成混合分量，(4)

p(x1:M) = Y

M

m=1

X

K

k=1

X

N

n=1 P

akak，n

i ai P j ai，j p(xm | k，n)。

(5)

在没有监督的情况下，通过最大化等式(5)中部分胶囊受稀疏性约束的可能性来训练模型，参见第2.4节。部分胶囊m可以被分配给目标胶囊k？作为k？= arg maxk akak，n p(xm | k，n)。实验结果表明，该模型能够对属于不同星座的点进行无监督的实例级分割，即使是在人类难以理解的数据中。参见图2的例子和3.1节的细节。

2.2零件胶囊自动编码器(PCAE)

将图像解释为零件的几何排列需要首先推断图像由哪些零件组成，以及零件与观察者的关系(我们称之为它们的姿态)。对CCAE来说，一个零件只是一个2D点，但在这里每个零件胶囊都有一个六自由度姿态，一个存在变量和一个独特的身份。我们将零件发现问题定义为自动编码:编码器学习推断不同零件胶囊的姿态和存在，而解码器学习每个零件的图像模板(图3)，类似于Tieleman，2014；Eslami等人，2016年。对应于当前部分的模板使用它们的姿态被仿射变换，并且这些变换模板的像素被用于为每个图像像素创建单独的混合模型。PCAE之后是一个物体胶囊自动编码器(OCAE)，它紧密地重新定位CCAE，详见第2.3节。

让y ∈ [0，1]h×w×c为图像。我们将部分胶囊的最大数量限制为M，并使用编码器推断它们的姿态xm ∈ R、存在概率dm ∈ [0，1]和特殊特征zm ∈ R cz，每部分胶囊一个。后者不参与直接图像重建，而是向OCAE通报相应部分的特殊方面；他们是通过从OCAE反向传播衍生品来训练的。

目前，我们不允许同一类型的零件在图像中多次出现，因此零件胶囊本身不会跨空间复制，尽管它们可以。然而，我们确实需要识别图像中出现的任何部分，因此编码器由一个具有自下而上注意机制的有线电视新闻网组成；对于胶囊k的每个部分，它预测具有空间维度he × we的6(姿态)+ 1(存在)+ cz(特殊特征)胶囊参数的特征图e k，以及单通道注意力掩模ak。胶囊的最终参数被计算为P i P j ek，I，j softmax(a)k，I，j，其中softmax沿着空间维度。这类似于全局平均池，但允许一些空间位置对最终结果的贡献大于

从胶囊特征向量导出这些矩阵允许可变形物体。我们将OPs建模为依赖于输入的分量和恒定偏差的总和。我们鼓励不同的胶囊专门针对不同的星座，对前者施加强大的L2惩罚。

我们将零件视为独立零件，并在相同的混合模型下评估它们的概率。虽然零件和预测之间没有明确的1:1联系，但它在实践中似乎运行良好。

3

图3:在MNIST(左)以及索贝尔滤波的SVHN(中)和CIFAR10(右)上学习的模板。在每种情况下，模板都收敛到笔画。对于SVHN，它们通常采取双笔画的形式——这是由于sobel滤波，它有效地提取边缘。

(a) (b) (c) (d) (e)

图4: 40 × 40 MNIST (a)图像及其(b)红色部分胶囊和绿色目标胶囊的重建，黄色重叠区域。对于每个输入(c)，先验地(左)只激活少数对象胶囊，而重构它(右)则需要更少的对象胶囊。最活跃的胶囊(d)捕捉对象身份和关于其外观的大部分信息。最后，(e)仿射变换模板显示了如何准确地使用零件来重建图像。

其他；我们称这种方法为基于注意力的汇集。第3.3节分析了其对模型性能的影响。

图像像素被建模为独立的高斯混合。对于每一个像素，我们取变换后的模板的相应像素，并将其视为具有恒定方差的各向同性高斯分量的中心。它们的混合概率与部分胶囊的存在概率和该位置颜色值的函数fc : R c 7→ [0，1]成比例，其中c是图像通道的数量。更正式地说:

x1:M，d1:M，z1:M = henc(y)将图像编码为部分胶囊参数，(6) Tbm = TransformImage(Tm，xm)将仿射变换应用于图像模板，(7) p y m，I，j ∝ dm fc Tbm，I，j计算混合概率，(8)

p(y) = Y

I，j

X

M

m=1

p y m，I，j N yi，j | Tbm，I，j，σy

2.3目标胶囊自动编码器(OCAE)

计算图像可能性。(9)

下一步是在已经发现的部分中找到对象。为此，我们使用级联姿态xm、特殊特征zm和展平模板Tm(传达零件胶囊的身份)作为OCAE的输入，这在以下方面不同于CCAE。首先，我们将部分胶囊存在概率dm输入OCAE编码器——这些概率用于偏置设置转换器的注意机制，以不考虑缺失点。其次，dm也被用来衡量部分胶囊的对数似然性，参见等式(5)。此外，我们停止梯度在所有OCAE的输入，除了特殊的特征，以提高训练稳定性和避免潜在变量崩溃的问题；例如，见Rasmus等人，2015年。最后，PCAE发现的部分具有独立的身份(模板和特殊特征，而不是2D点)。因此，每个部分姿态被解释为来自对象胶囊的预测的独立混合物——其中每个对象胶囊精确地做出M个候选预测Vk、1:M，或者每个部分精确地做出一个候选预测。因此，部分胶囊可能性由下式给出:

p ( x1: M，d1: M) = YMm=1 " XKk=1 P a k ak，m i a i P j ai，j p ( x m | k，m ) # d m

(10)

。

2.4实现稀疏多样的胶囊存在

堆叠胶囊自动编码器被训练成最大化像素和零件对数似然性(Lll =对数p(y) +对数p(x1:M))。然而，如果不受约束，它们倾向于使用所有的零件和对象胶囊来解释每个数据示例，或者折叠到使用胶囊的始终相同的子集，而不管输入。我们希望该模型针对不同的输入示例使用不同的部件封装集，并将对象封装专门用于部件的特定排列；

模板被认为是稀疏的；如果在给定位置存在非零值的模板，则应该使用该模板。

在推理过程中，发现的对象不会自上而下地用于细化零件的存在或姿态。然而，通过OCAE反向传播的导数改进了推断零件的低级编码器网络。

4

目标胶囊

CNN

模板

设置

变压器

特色

出席

图像

可能性

转换模板

部分

可能性

信息流。梯度反向流动。

梯度停止的信息流。

可训练变量激活损失

稀少

部分胶囊

摆出存在的姿势

对象-查看器关系

固定对象-零件关系

预测零件姿态

图5: SCAE建筑。

为了鼓励这一点，我们强加稀疏性和熵约束。我们在第3.3节中评估它们的重要性。

我们首先定义前面和后面的物体-胶囊存在如下。对于具有K个目标胶囊和M个部分胶囊的大小为B的小批次，我们定义了先前胶囊存在的小批次，先前1:K，尺寸为[，K]，而后胶囊存在的小批次，后部1:K，1:M，尺寸为[，B，K，M]，

a前k = ak最大m am，k，a后k，m = akak，m N (xm | m，k)，(11)

分别地；前者是来自对象胶囊k的预测中的最大存在概率，而后者是用于解释部分胶囊m的非标准化混合概率。先验稀疏性让uk = B PB = 1a先验B，k是不同训练示例中对象胶囊k的平均存在概率，并且ubb = PK k=1 a先验B，k是给定示例中对象胶囊存在概率的总和。如果我们假设训练示例随机均匀地包含来自不同类别的对象，并且我们希望为每个类别分配相同数量的对象胶囊，那么每个类别将获得K/C胶囊。此外，如果我们假设每个图像中只存在一个对象，那么对于每个输入示例，应该存在B/C对象胶囊。为此，我们尽量减少，

lprior = B×B

b=1

| | ubb kC | | 2+kX K

k=1

||英国B C ||2。(12)

后稀疏相似性，让vk和vbb分别是P k，m a后b，k，m和P b，m a后b，k，m的归一化版本。我们发现最小化胶囊后部存在的实例内熵(vk)和最大化其实例间熵(vbb)是有益的，其中H是熵。最终的损失是，

Lposterior = K X K

k=1

h(vk)B×B

b=1

H(vbb)。(13)

每个活动的目标胶囊应该解释至少两个部分。我们说，如果在其他目标胶囊中，目标胶囊具有该部分的最高后验混合概率，则该目标胶囊已经“赢得”了该部分。然后我们为每个对象胶囊创建二进制标签，如果胶囊赢得至少两个部分，标签为1，否则为0。最终的损失采取生成的标签和先前胶囊存在之间的二元交叉熵的形式。这种损失仅用于点数据上的独立星座模型实验，参见第2.1节和第3.1节。

图5示出了SCAE的示意性架构。我们优化图像和部分可能性以及辅助损失的加权和。附录一解释了失重选择过程以及用于实验的数值

为了使存在概率值(ak、ak、m和dm)更接近二进制，我们将均匀噪声∈[2，2]注入逻辑，类似于Tieleman，2014。这迫使模型预测远离零的逻辑，以避免随机性，并使预测的存在概率接近二进制。有趣的是，在我们的案例中，它往往比使用具体分布更有效(马迪森等人，2017年)。

3评估

SCAE的解码器使用明确参数化的仿射变换，允许编码器的输入用一小组变换的对象或部分来解释。以下评估显示

5

表1:无监督分类结果(标准偏差)的百分比在5次运行中平均。基于互信息的方法用阴影表示。标有“使用数据增强”的结果，∇使用经图像预处理的特征代替图像，同时取自Ji等人，2018年。根据双边t检验，我们强调最好的结果和那些在98%置信区间内的结果。

方法MNIST投药10号

KMEANS (Haeusser等人，2018年)53.49 20.8 12.5

AE (Bengio等人，2007年)81.2 31.4 -

甘(拉德福德等人，2016年)82.8 31.5 -

IMSAT (Hu等人，2017)†∇98.4(0.4)45.6(0.8)57.3(3.9)

IIC (Ji等人，2018年)98.4 (0.6) 57.6 (5.0)

模数转换器(Haeusser等人，2018年)98.7(0.6)29.3(1.5)38.6(4.1)MAXACT(SCAE)98.0(. 15)19.79(1.0)49.07(1.7)CLUSTNN(SCAE)98.5(. 11)19.39(1.5)53.0(3.8)LINMATCH(SCAE)98.5(. 10)25.01

嵌入的几何知识如何帮助发现数据中的模式。首先，我们展示了CCAE在由二维点构成的星座排列中发现底层结构，从而执行实例级分割。其次，我们将OCAE和PCAE配对，并调查由此产生的SCAE是否能在真实图像中发现结构。最后，我们提出了消融研究，表明模型的哪些组成部分对结果有贡献。

3.1发现星座

我们在线创建星座排列，其中每个输入示例由多达11个属于多达三个不同星座(两个正方形和一个三角形)的二维点以及指示点存在的二元变量(点可能会丢失)组成。每个星座的概率为0.5，并经历相似性变换，由此随机缩放、旋转高达180°并移位。最后，对每个输入示例进行归一化，使得所有点都位于[1，1]内。请注意，我们使用点集而不是图像作为模型的输入。我们将CCAE与使用相同编码器但解码器更简单的基线进行比较:解码器使用胶囊参数向量ck来直接预测四个点的位置、精度和存在概率以及整个对应星座的存在概率。实施细节见附录a1

两种模型都是通过最大化零件对数似然性来进行无监督训练的。我们通过尝试将每个输入点分配给一个对象胶囊来评估它们。为此，我们将每个输入点分配给该点后验概率最高的目标胶囊，参见第2.1节，并计算分割精度(即真阳性率)。

CCAE始终达到4%以下的误差，最佳模型达到2.8%，而使用相同的超参数搜索预算，最佳基线达到26%的误差。这表明，对几何关系建模的感应偏置接线有助于将误差降低一个数量级——至少在玩具装置中，每组点由已独立变换的熟悉星座组成。

3.2无监督的类发现

为了允许特定类的对象出现多模态，我们通常使用比类标签数量更多的对象胶囊。我们期望目标胶囊存在概率的向量应该是类标签的高度信息。为了检验这一假设，我们在MNIST、西佛恩和西佛10上训练SCAE，并试图给目标胶囊存在的向量分配类别标签。这是通过以下方法之一来完成的:MAXACT:我们搜索一个最大限度地激活给定对象胶囊并为此胶囊分配相应标签的训练示例；聚类神经网络(CLUSTERNN):我们将KMEANS聚类成C聚类，然后找到最接近每个聚类质心的训练样本，给聚类分配一个标签；LINMATCH:在用KMEANS找到10个聚类后，我们使用二分图匹配(Kuhn，1955)来找到将分类误差最小化的聚类索引排列——这是无监督分类的标准实践，例如参见Ji等人，2018；LINPRED:我们在给定存在向量的情况下训练一个具有监督的线性分类器；这学习了K × 10权重和10偏差，其中K是对象胶囊的数量，但是它不修改主模型的任何参数。

所有考虑的数据集都有10个类。

6

表2:MNIST消融研究。所有使用的模型组件都有助于其最终性能。AFFNIST的结果显示了非分布通用化特性，并且来自于在40 × 40 MNIST上训练的模型。数字代表10次运行的平均值%和(标准偏差)。根据双边t检验，我们强调最好的结果和那些在98%置信区间内的结果。

方法MNIST 40 × 40 MNIST空军司令部

完整型号97.0 (.87) 98.5 (.1) 92.2 (.59)

无后稀疏度96.7 (.7) 98.2 (.48) 87.6 (1.63)

a)没有先前的稀疏度90.5 (7.56) 94.0 (3.03) 74.0 (4.94)

无前后稀疏度63.0(13.48)62.7(10.46)40.7(6.81)物体盖96.4 (1.41) 97.8 (.67) 90.8 (2.97) b)任何盖84.8(6.22)85.1(13.13)76.3(12.89)零件盖83.9(7.57)83.9(7.57)无噪声

无变形87.6 (6.13) 95.2 (1.04) 87.6 (1.26)

d)线性零件enc 94.8 (3.0) 98.1 (.26) 76.3 (2.22)

CONV部分enc 96.3 (.85) 97.8 (.95) 80.1 (2.58)

e)物体帽的MLP enc 73.0(6.34)70.3(11.2)52.5(11.29)

f)无特殊功能63.1(10.55)66.9(23.59)50.5(18.26)

与以前关于无监督聚类的工作一致(Ji等人，2018年；Hu等人，2017年；Hjelm等人，2019年；Haeusser等人，2018年)，我们训练我们的模型并在完整数据集上报告结果(训练、有效和测试分割)。我们方法的LINPRED变体中使用的线性变换是在各个数据集的TRAIN分割上训练的，同时报告了它在测试分割上的性能。我们使用了PCAE，MNIST有24个单通道11 × 11模板，SVHN和CIFAR10分别有24个和32个三通道14 × 14模板。我们使用sobel滤波图像作为SVHN和CIFAR10的重建目标，如Jaiswal等人，2018年，同时使用原始像素强度作为PCAE的输入。OCAE分别使用了24、32和64个目标胶囊。关于模型架构和超参数调整的更多细节见附录A。所有结果见表1。SCAE在MNIST和SVHN上的无监督对象分类中取得了有竞争力的结果，在CIFAR10上的表现稍差，这将在第5节中进一步讨论。

3.3消融研究

SCA有许多活动部件；消融研究显示了哪些模型组件是重要的以及在何种程度上是重要的。我们在MNIST训练SCAE变体以及填充和翻译的40 × 40版本数据集，其中原始数字在每个方向上被翻译多达6个像素。在两个数据集的测试分割上测试训练的模型；此外，我们对在40 × 40 MNIST上训练的模型进行了评估。在AFFNIST上的测试表明该模型是否可以推广到看不见的观点。罗林森等人在2018年使用该任务评估稀疏无监督胶囊，准确率达到90.12%。SCAE达到92.2±0.59%，表明其观点泛化能力较好。我们选择LINMATCH性能指标，因为它是无监督分类社区喜欢的指标。结果分为几组，如表2所示。我们依次描述每一组。组a)表明，第2.4节中引入的稀疏性损失提高了模型性能，但后验损失可能不是必需的。b组)检查向逻辑中注入噪声对存在概率的影响，参见第2.4节。将噪声注入部分胶囊似乎至关重要，而目标胶囊中的噪声似乎没有必要——后者可能是由于稀疏性损失。c组)显示，在解码器中使用相似性(相对于仿射)变换在某些情况下可能是限制性的，而不允许变形在任何情况下都会损害性能。

d组)评估零件胶囊编码器的类型。线性编码器需要一个有线电视新闻网和一个完全连接的层，而CONV编码器为每个胶囊参数预测一个特征图，然后是全局平均池。部分胶囊编码器的选择似乎对内部分配性能没有多大影响；然而，当在不同的数据集上进行评估时，我们基于注意力的汇集确实实现了更高的分类精度，显示出对新观点的更好概括。

此外，使用设置转换器作为对象胶囊编码器是必要的。我们假设这是由于设置转换器寻找集群的自然趋势，正如李等人在2019年所报告的。最后，f)使用特殊特征zm似乎并不逊色——大概是由于高级胶囊对主编码器所学习的表示的影响。

7

4相关工作

胶囊网络我们的工作结合了转换自动编码器(辛顿，克里哲夫斯基等人，2011年)和电磁胶囊(辛顿，萨布尔等人，2018年)的想法。通过训练自动编码器从原始图像加上明确表示变换的额外输入来预测输入图像的仿射变换版本，变换自动编码器消除了仿射感知胶囊实例化参数。相比之下，我们的模型除了图像不需要任何输入。电磁胶囊和前面的动态胶囊(Sabour等人，2017)都使用零件的姿态和学习的零件→对象关系来投票选择对象的姿态。当多个部分投出非常相似的票时，该对象被假定存在，这通过交互式推理(路由)算法来实现。迭代路由效率很低，并推动了进一步的研究。王和刘，2018年将路由表述为聚类损失和基于KL散度的正则项的优化。Zhang等人2018年提出了一种基于加权核密度估计的路由方法。Li等人，2018年提出用两个分支近似路由，并通过两个分布(较低和较高的胶囊)之间的最佳传输发散发送反馈。与以前的工作相比，我们使用对象来预测零件，而不是相反，因此我们可以在推理时省去迭代路由。OCAE的编码器学习如何将零件分组到对象中，并且它尊重单亲约束，因为它是使用解码器产生的导数来训练的，解码器使用零件的混合模型，假设每个零件必须由单个对象来解释。

此外，由于预测零件的是对象，零件在姿态上的自由度比对象(如在CCAE)要少。推断仍然是可能的，因为OCAE编码器基于所有部分而不是单个部分进行对象预测。我们这种胶囊的另一个优点是它可以进行无监督学习。先前版本的胶囊使用了区别性学习，尽管罗林森等人，2018年使用了Sabour等人，2017年引入的重构MLP，在没有监督的情况下训练动态胶囊，并且已经表明胶囊条件重构的无监督训练有助于对AFFNIST分类的推广；我们进一步改进了他们的结果，参见第3.3节。

无监督分类在计算机视觉中有两种主要的无监督目标类别检测方法。第一种是基于表示学习，通常需要在学习到的表示之上发现聚类或学习分类器。Eslami等人，2016年；Kosiorek等人，2018年使用迭代过程来推断可变数量的潜在变量，场景中的每个对象对应一个潜在变量，这些潜在变量对对象类具有很高的信息量，而Greff等人，2019年；Burgess等人，2019以迭代方式执行无监督的实例级分割。虽然与我们的工作相似，但这些方法不能将对象分解成它们的组成部分，也不能提供对象形状的明确描述(例如，模板及其在我们模型中的姿态)。第二种方法通过最小化基于互信息的损失和直接学习类别分配概率来明确目标分类。IIC (Ji等人，2018年)最大化了描述输入图像的(变换的)版本的两个离散概率向量之间的MI的精确估计。深度信息最大值(Hjelm等人，2019年)依赖于负样本，并通过噪声对比估计最大化预测概率向量与其输入之间的最小干扰(Gutmann和Hyvä rinen，2010年)。这类方法直接最大化分配给离散集群的信息量，并且在大多数无监督的分类任务中拥有最先进的结果。基于多元智能的方法具有互信息估计的典型缺点:它们需要大量的数据扩充和大批量。这与我们的方法形成对比，我们的方法在批量不大于128并且没有数据增加的情况下实现了可比较的性能。几何推理将几何知识纳入神经网络的其他尝试包括利用群变换的等方差性质(科恩和韦林，2016年)或新型卷积滤波器(奥亚隆和马拉特，2015年；Dieleman等人，2016年)。尽管与标准CNNs相比，它们在处理旋转或反射时实现了显著的参数效率，但这些方法不能处理仿射变换的额外自由度，如尺度。Lenssen等人，2018年将胶囊网络与群卷积相结合，以保证胶囊网络的等变和不变性。空间转换器；贾德伯格等人，2015)将仿射变换应用于图像采样网格，同时转向网络(科恩和韦林，2017；雅各布森等人，2017)动态改变卷积滤波器。这些方法与我们的相似之处在于，变换参数是由神经网络预测的，但不同之处在于，ST使用应用于整个图像的全局变换，而可控网络使用

8

只有局部转换。我们的方法可以对每个对象使用不同的全局转换，也可以对它们的每个部分使用局部转换。

5讨论

我们工作的主要贡献是一种新颖的表示学习方法，其中使用高度结构化的解码器网络来训练一个可以将图像分割成部分及其姿态的编码器网络，以及另一个可以将部分组成连贯整体的编码器网络。尽管我们的训练目标与分类或聚类无关，但SCAE是在无监督对象分类中获得竞争结果而不依赖于互信息的唯一方法。这很重要，因为与我们的方法不同，基于管理信息的方法需要复杂的数据扩充。通过使用基于心肌梗塞的损失来训练SCAE，可能进一步改善结果，在那里胶囊概率向量可以在IIC扮演离散概率向量的角色(Ji等人，2018年)。SCAE在CIFAR10上表现不佳，这可能是因为使用了固定模板，这些模板的表达能力不足以模拟真实数据。这可以通过构建胶囊自动编码器的更深层次(例如，计算机图形中的复杂场景被建模为仿射变换几何图元的深树)以及使用依赖于输入的形状函数而不是固定模板来解决，这两者都是未来工作的有希望的方向。还可以通过在生成模型中使用可微分渲染器来制作更好的PCAE，用于学习主胶囊，该可微分渲染器从主胶囊重建像素。

最后，SCAE可以是混合模型的“图形”部分，它还包括一个通用的“地面”部分，可以用来说明除图形以外的一切。然后，可以使用顺序注意力一次感知一个图形来分析复杂的图像。

6致谢

我们要感谢黄珊迪在编辑手稿和制作图1方面的帮助。此外，我们要感谢阿里·伊斯拉米和丹尼尔·哈夫纳在整个项目中的有益讨论。我们也感谢金现代、马丁·恩格尔克、埃米莉恩·杜邦和西蒙·科恩布利对初稿的反馈。

参考

J.文学士、基罗斯和辛顿(2016年)。“层标准化”。In: CoRR. arXiv: 1607.06450。

Y.本吉奥、朗布林、波博维奇和拉罗彻尔(2007年)。“深度网络的贪婪分层训练”。神经信息处理系统的进展。

C.P. Burgess，L. Matthey，N. Watters，R. Kabra，I. Higgins，M. Botvinick和A. Lerchner (2019年)。

“莫奈:无监督场景分解和表示”。In: CoRR. arXiv: 1901.11390。科恩和韦林(2016)。“群等变卷积网络”。国际机器学习会议。

T.科恩和韦林(2017)。“可控有线电视新闻网”。国际表征学习会议。

S.介曼、德法乌和卡武科格鲁(2016)。“在卷积神经网络中利用循环对称性”。In: CoRR. arXiv: 1602.02660。

S.M. A. Eslami、N. Heess、T. Weber、Y. Tassa、D. Szepesvari、K. Kavukcuoglu和G. E. Hinton

(2016年)。“注意、推断、重复:用生成模型快速理解场景”。神经信息处理系统的进展。arXiv: 1603.08575。

K.格里夫、考夫曼、卡巴拉、沃特斯、伯吉斯、佐兰、马特西、博蒂尼克和

A.勒施纳(2019)。“迭代变分推理的多目标表示学习”。In: arXiv预印本arXiv:1903.00450。

M.Gutmann和a . Hyv rinen(2010年)。“噪声对比估计:一种新的估计原则

对于非标准化的统计模型”。国际人工智能和统计会议。

9

P.Haeusser，J. Plapp，V. Golkov，E. Aljalbout和D. Cremers (2018年)。“关联深度聚类:

培训无标签的分类网络”。德国模式识别会议。韩丁(1979)。”心理意象中结构描述效果的一些演示”。In:认知科学3。

G.韩丁、克里哲夫斯基和王世德(2011)。“转换自动编码器”。国际人工神经网络会议。

G.E. Hinton、S. Sabour和N. Frosst (2018年)。“带电磁路由的矩阵胶囊”。国际学习表征会议。

R.哈耶姆、费多罗夫、拉瓦伊-马奇尔登、格雷瓦尔、特里施勒和本吉奥(2019年)。

“通过互信息估计和最大化学习深度表征”。In: CoRR. arXiv: 1808.06670。

W.Hu，T. Miyato，S. Tokui，E. Matsumoto和M. Sugiyama (2017)。“学习离散表示——

通过信息最大化自我增强训练的陈述”。国际机器学习会议。

J.H. Jacobsen、B. De Brabandere和a . w . Smelders(2017年)。“深层剩余网络中的动态可控区块”。In: CoRR. arXiv: 1706.00598。

M.Jaderberg，K. Simonyan，A. Zisserman和K. Kavukcuoglu (2015年)。“空间转换器

网络”。神经信息处理系统的进展。DOI:10.1038/NBT . 3343 . ARXiv:1506.02025 v1。

A.Jaiswal、W. AbdAlmageed、Y. Wu和P. Natarajan (2018年)。“胶囊:生殖对抗胶囊网络”。在:欧洲计算机视觉会议(ECCV)。X. Ji、J. F. Henriques和A. Vedaldi (2018年)。“无监督的不变信息蒸馏

图像分割和聚类”。In: CoRR. arXiv: 1807.06653。网址:http://arxiv.org/ ABS/1807 . 06653。

A.Kosiorek，H. Kim，Y. W. Teh和I. Posner (2018年)。“顺序出席、推断、重复:生成

运动物体的建模”。神经信息处理系统的进展。arXiv: 1806.01794。

H.W. Kuhn (1955)。“分配问题的匈牙利方法”。In:海军研究后勤季刊。

J.李、李、金、科西莱克、崔世杰和崔永伟(2019)。“设置变压器”。国际机器学习会议。arXiv: 1810.00825。

J.E. Lenssen、M. Fey和P. Libuschewski (2018年)。“集团等变胶囊网络”。神经信息处理系统的进展。

H.李晓国、戴秉国、欧阳w和王晓明(2018)。“神经网络封装”。In: CoRR. arXiv: 1808.03749。

C.马迪森、阿姆尼和蒂赫(2017年)。“具体分布:离散随机变量的连续松弛”。国际学习表征会议。E. Oyallon和S. Mallat (2015年)。“物体分类的深旋转平移散射”。美国电气与电子工程师协会计算机视觉和模式识别会议。

A.拉德福德、梅斯和钦塔拉(2016年)。“深度无监督表示学习

卷积生成对抗网络”。国际学习表征会议。

A.拉斯穆斯、贝里隆德、洪卡拉、瓦尔波拉和赖科(2015年)。“梯形网络的半监督学习”。神经信息处理系统的进展。

D.罗林森、艾哈迈德和科瓦德洛(2018年)。“稀疏无监督胶囊更好地概括”。In: CoRR. arXiv: 1804.06094。

I. Rock (1973年)。方向和形式。学术出版社。

S.萨布尔、弗罗斯特和辛顿(2017)。“胶囊之间的动态路由”。神经信息处理系统的进展。

T.Tieleman和G. E. Hinton (2012年)。讲座6.5——均方根过程:用梯度除以其最近量值的运行平均值。用于机器学习的神经网络。

10

T.Tieleman (2014)。优化生成图像的神经网络。加拿大多伦多大学。

D.王和刘问(2018)。“胶囊之间动态路由的优化视图”。作者:张学良、周清和吴晓明(2018)。“基于加权核密度的快速动态路由

估计”。国际人工智能和机器人研讨会。

11

模型细节

a1星座实验

CCAE使用四层集变压器作为编码器。每层有四个注意头，每个头有128个隐藏单位，然后是层规范(Ba等人，2016年)。编码器输出三个32维向量——每个对象胶囊一个向量。解码器对每个对象胶囊使用单独的神经网络来预测用于建模其点的所有参数:这包括对总共12个候选胶囊的四个候选部分预测。在这个实验中，每个物体→零件关系OP只是物体参照系中的一个二维偏移(而不是一个3 × 3的矩阵)，它被相应的OV矩阵仿射变换以预测二维点。

a2图像实验

我们对部分胶囊使用卷积编码器，对目标胶囊使用集合转换器编码器(Lee等人，2019年)。从对象胶囊到部分胶囊的解码是用MLPs完成的，而输入图像是用仿射变换的学习模板重建的。表3提供了我们使用的体系结构的详细信息。

表3:架构细节。s在最后一列中表示条目与SVHN相同。数据集MNIST星群SVHN CIFAR10号模板编号:A 24 24 32模板大小:n11×11 14×14S号胶囊3 24 32 64部分:有线电视新闻网编号:2x(128:2)-2x(128:1)-2x(128:1)-2x(128:2)S组变压器4x(4-128)-32 3x(1-16)-256 3x(2-64)-128S

我们使用ReLu非线性，除了存在概率，我们使用sigmoids。(128:2)对于有线电视新闻网意味着128个频道，步幅为2。所有内核都是3 × 3。对于设置变压器(1-16)-256表示一个注意头、16个隐藏单元和256个输出单元；它使用层标准化(Ba等人，2016年)，与原始论文(Lee等人，2019年)一样，但没有丢失。所有实验(除星座外)每部分胶囊使用16个特殊特征。

对于SVHN和CIFAR10，我们使用归一化的sobel滤波图像作为重建的目标，以强调形状的重要性。附录B中的图6显示了SVHN和CIFAR10重建的示例。滤波过程如下:1)应用索贝尔滤波，2)减去中值颜色，3)取图像的绝对值，4)归一化图像值为∈ [0，1]。所有模型都经过RMSProp优化器(Tieleman和Hinton，2012)动量= .9和= (10 \* batch\_size)的训练。星座的批量为64，所有其他数据集为128。对于MNIST和星座实验，学习率等于10(没有任何衰减)，而我们对SVHN和CIFAR10进行超参数搜索:我们搜索的学习率在5 \* 10到5 \* 10的范围内，并且每10或3 \* 10次权重更新，指数学习率衰减0.96。SVHN和CIFAR10的学习率均为10，SVHN的衰减步长为10，CIFAR10的衰减步长为3 \* 10。验证集中的LINPRED精度用作选择最佳超参数的代理，包括不同损失的权重，如表4所示。模型在单个特斯拉V100图形处理器上最多经过3 \* 10次迭代训练，星座实验耗时40分钟，CIFAR10耗时不到一天。

表4:损失权重值。稀疏性损失中的量词内部和之间对应于等式(12)和(13)的不同项。

数据集星座MNIST卫星星座

部分ll重量1 1 2.56 2.075

图像ll权重不适用1 1 1

稀疏度1 1 0.22 0.17内的先验

稀疏度1 1 0.1 0.1之间的先验

稀疏度0 10 8.62 1.39内的后部

稀疏度0 10 0.26 7.32之间的后部

太少活性胶囊10 0 0 0

12

重建

SVHN

CIFAR10

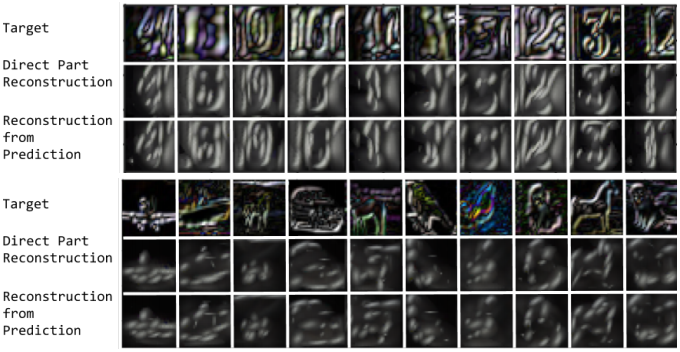


图6: 10样本SVHN和Cifar10重建。第一行显示索贝尔滤波后的目标图像。第二行显示了从零件胶囊层直接重建。如果我们使用零件姿态的对象预测而不是零件姿态本身来进行重建，则第三行显示重建。该模型中的模板与图像具有相同数量的通道，但是它们已经收敛到黑白模板，并且重建没有颜色多样性。SCAE模型完全是在无监督的情况下训练的，但是重建往往集中在奇异值分解的中心数字上，并过滤掉剩余的杂波。

13