Alibaba Group 阿里巴巴集团 | 阿里技术

阿里巴巴机器智能

计算机视觉技术精选

顶级学术会议 CVPR-2018 收录论文

| 卷积神经网络
| 生成对抗网络
| 零样本学习
| 跨模态检索

Alibaba Group
阿里巴巴集团

阿里技术

扫一扫二维码图案，关注我吧

「阿里技术」微信公众号

「阿里巴巴机器智能」微信公众号

| 序言

CVPR（Conference on Computer Vision and Pattern Recognition）是计算机视觉领域的顶会之一，伴随着视觉 AI 的火热，近几年参会人数急剧增加。2018 年的 CVPR 会议于 6 月 18 日-22 日在美国犹他州盐湖城举办。本届大会有超过 3300 篇的大会论文投稿，录取 979 篇（接受率约为 29%），其中包括 70 篇 Oral 论文和 224 篇 Spotlight 论文，参会人数达到 6500 人。除了正会以外，本届 CVPR 有 21 个 tutorials 和 48 个 workshops，以及超过 115 个公司的工业展会。

近些年伴随着深度学习技术、GPU 和云计算等运算力的增强，计算机视觉技术进入越来越实用的阶段。无论是在电商、安防、娱乐，还是在工业、医疗、自动驾驶，计算机视觉技术都扮演着越发重要的角色。在阿里巴巴广阔的商业和数据生态的发展中，计算机视觉技术的研发和商业化落地密不可分。比如拍立淘利用图像搜索和识别技术帮助淘宝、天猫、AliExpress, Lazada 等电商 app 的用户在移动端通过拍照就能找到相同相似的商品，从而进行更加方便的购物。比如在线下新零售领域，阿里研发了人的追踪和空间定位、货架商品 SKU 识别等技术去推动商场、超市、酒店等的人货场数字化，并在此基础上做进一步的商业分析。在城市大脑项目，阿里研发了大规模视频高效处理，人和车辆的搜索和识别等技术帮助城市交通事故识别，人流轨迹判断以及交通数据样本汇总。

在本届 CVPR 顶会中，阿里巴巴总共发表 18 篇论文。此外，阿里巴巴也举办了展台展示、学者晚宴、展台技术 Talk 等多项活动，把包括图像搜索、城市大脑、自动驾驶、FashionAI、鹿班设计、三维物体建模、交互仿真虚拟人、广告、多媒体智能审核等阿里巴巴在 CV 领域的技术成果和应用情况集中亮相国际舞台。在这本论文合集中，我们收录了其中有代表性的 7 篇论文。

Spotlight 论文《基于时间尺度选择的在线行为预测》讨论了视频中行为预测的一个非常重要的问题：怎么去选择一个好的时间维度窗口？论文提出了多个子网络的尺度选择网，包括时间序列建模的一维卷积子网络，尺度回归子网络，以及行为预测子网络。在两个公开数据集上，尺度选择网的实验结果优于其他方法，并且准确率也接近使用 Ground Truth 尺度的结果。

Spotlight 论文《基于语境对比特征和门控多尺度融合的场景分割》致力于场景分割中的两大问题：场景图片中像素形式的多样化（例如，显著或者不显著，前景或者背景）和场景图片中物体大小的多样性。文章针对这两个问题分别提出了语境对比局部特征和门控多尺度融合方法。本文提出的模型在 Pascal Context, SUN-RGBD 和 COCO Stuff 三个场景分割数据集上验证了性能，取得了目前最高的场景分割性能。

对于跨模态检索而言，如何学到合适的特征表达非常关键。Spotlight 论文《所见所想所找－基于生成模型的跨模态检索》提出了一种基于生成模型的跨模态检索方法，该方法可以学习跨模态数据的高层次特征相似性，以及目标模态上的局部相似性。本文通过大量的实验证明了所提出

的方法可以准确地匹配图像和文本，并且在 MSCOCO 以及 Flickr30K 的数据集上都取得了 state-of-the-art 的效果。

在论文《整体还是局部？应用 Localized GAN 进行图像内容编辑、半监督训练和解决 mode collapse 问题》中，作者建立了 GAN 和半监督机器学习中 Laplace-Beltrami 算子的联系，在用少量标注样本训练深度学习模型上取得了优异的性能。同时论文还展示了用 Localized GAN (LGAN) 对给定图像在局部坐标系下进行编辑修改，从而获得具有不同角度、姿态和风格的新图像；以及如何从流型切向量独立性的角度来解释和解决 GAN 的 mode collapse 问题。

论文《处理多种退化类型的卷积超分辨率》针对现有基于 CNN 的单图超分(SISR)算法不能扩展到用单一模型解决多种不同的图像退化类型的问题，提出了一种维度拉伸策略，使得单个卷积超分辨率网络能够将 SISR 退化过程的两个关键因素（即模糊核和噪声水平）作为网络输入来解决这个问题。实验结果表明提出的卷积超分辨率网络可以快速、有效的处理多种图像退化类型，为 SISR 实际应用提供了一种高效、可扩展的解决方案。

论文《于尺度空间变换的本征图像分解》将把图像分解为其本征的反射图像和光照图像看作是一个图像到图像的转换问题，并且将输入和输出在尺度空间进行分解。通过将输出图像（反射图像和光照图像）扩展到它们的拉普拉斯金字塔的各个成分，论文提出了一种多通道网络结构，可以在每个通道内并行地学习到一个图像到图像转换函数，这个函数通过一个具有跳过连接的卷积神经网络来表示。在 MPI-Sintel 数据集和 MIT Intrinsic Images 数据集上结果表明，新提出的模型在比之前最先进的技术上有了明显的进步。

大多数现有的零样本学习（Zero-Shot Learning，ZSL）方法都存在强偏问题。在论文《基于直推式无偏嵌入的零样本学习》中，作者提出了一个简单而有效的方法，称为准完全监督学习（QFSL），来缓解此问题。假定标记的源图像和未标记的目标图像都可用于训练。在语义嵌入空间中，被标记的源图像被映射到由源类别指定的若干个嵌入点，并且未标记的目标图像被强制映射到由目标类别指定的其他点。在 AwA2，CUB 和 SUN 数据集上进行的实验表明，文章的方法在遵循广义 ZSL 设置的情况下比现有技术的方法优越。

当下计算机视觉技术无疑是 AI 浪潮中火热的题目，受关注的程度持续升温。视觉技术的渗透，既可能是对传统商业的改造使之看到新的商业机会，还可能是创造了全新的商业需求和市场。好的视觉技术不仅需要有好的方法指引，而且需要在实际的场景中形成数据闭环和不断打磨。未来的计算机视觉技术一定是理论探索和数据实践的共同推进。希望这本论文合集能抛砖引玉，给学术界和工业界带来一些输入，共同推进计算机视觉技术的发展。

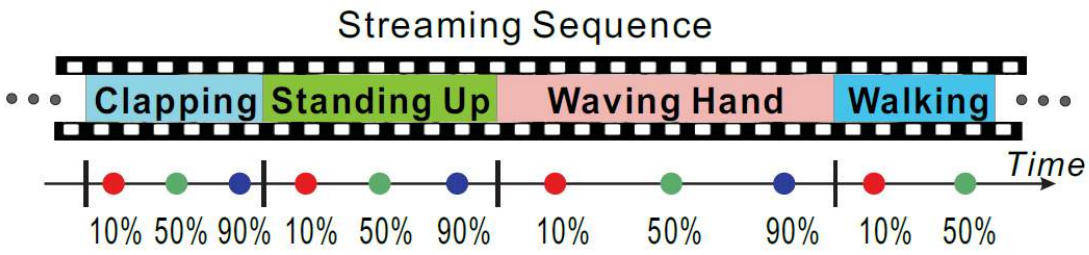
阿里巴巴资深算法专家 潘攀（启磐）

2018 年 12 月　于北京

目录

CVPR-18 阿里巴巴 Spotlight 论文：基于时间尺度选择的在线行为预测

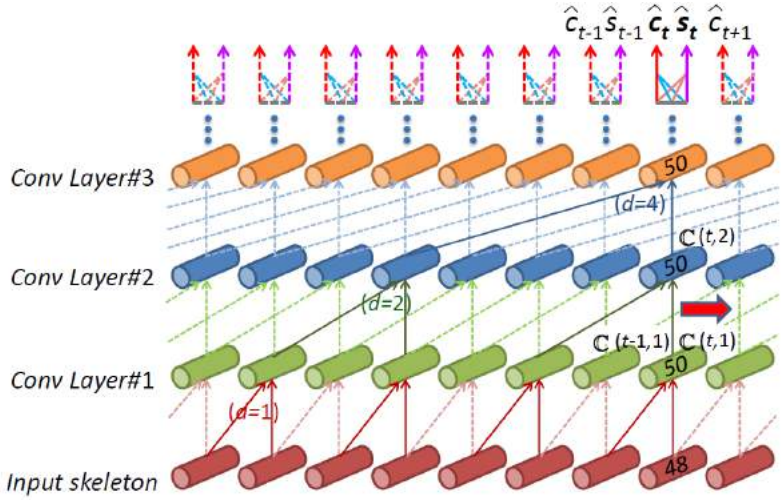
在线行为预测指的是当一个动作还未执行完之前，算法使用已经观测到的这些片段来预测该动作的类别。这个问题有几个关键点，首先，它是“在线”的，这表示算法得足够快，以实现在线应用；其次，算法需要在动作发生的早期（比如只完成了 10%）便进行类别预测；此外，算法处理的是未分割的视频，这意味着视频可能包含多个动作实例，比如下图的视频序列就包含了多个动作。



针对在线行为识别这一问题，我们可以使用在时间维度上滑动窗口的设计。传统滑动窗口方法往往采用一个固定的窗口尺度，或者采用多个尺度多次往返扫描。而在在线行为预测这一问题中，如果使用多次扫描的话，会影响算法的运行效率；但是如果只用单一固定尺度的话，选择一个合适的时间窗口尺度却很不容易。

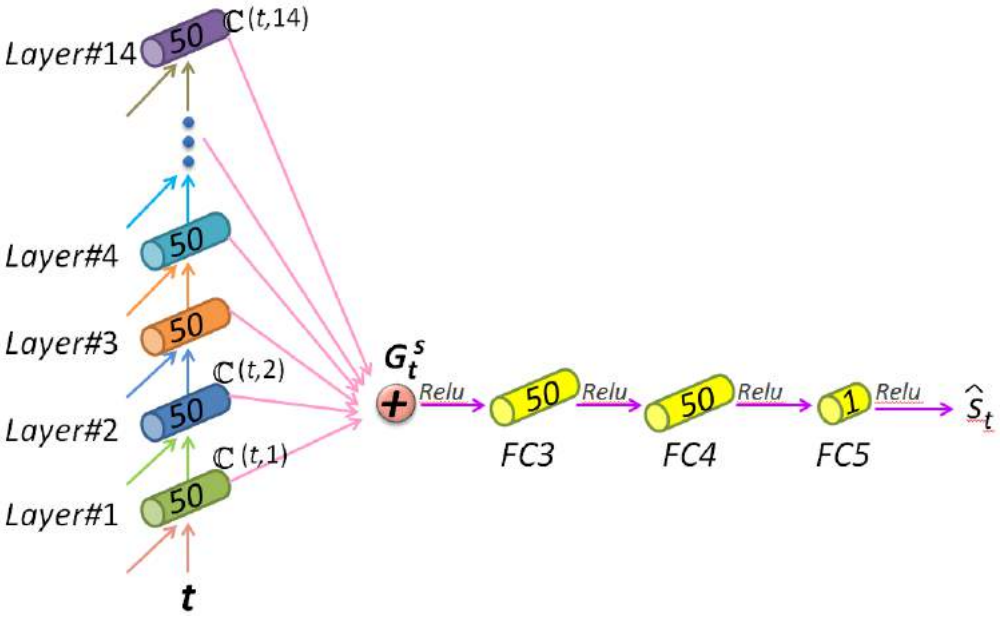
这是因为在行为预测任务中，当前正在发生的动作的已观测部分的长度在不同的时间点是在变化的。在动作发生的早期，我们需要使用比较小的时间窗口尺度，因为太大的窗口会包含很多来自于之前动作的帧，这些噪声信息会干扰对当前动作类别的识别。而在动作发生的后期，我们可以使用大的窗口尺度来尽可能覆盖该动作已执行的片段，以达到更好的预测准确率。这意味着在不同阶段使用一个固定的窗口尺度是不合适的。

在这篇论文中，作者提出了一个“尺度选择网”（scale selection network）来在不同时刻点动态的选择当前最合适的窗口尺度。该网络的基本结构如下图所示。



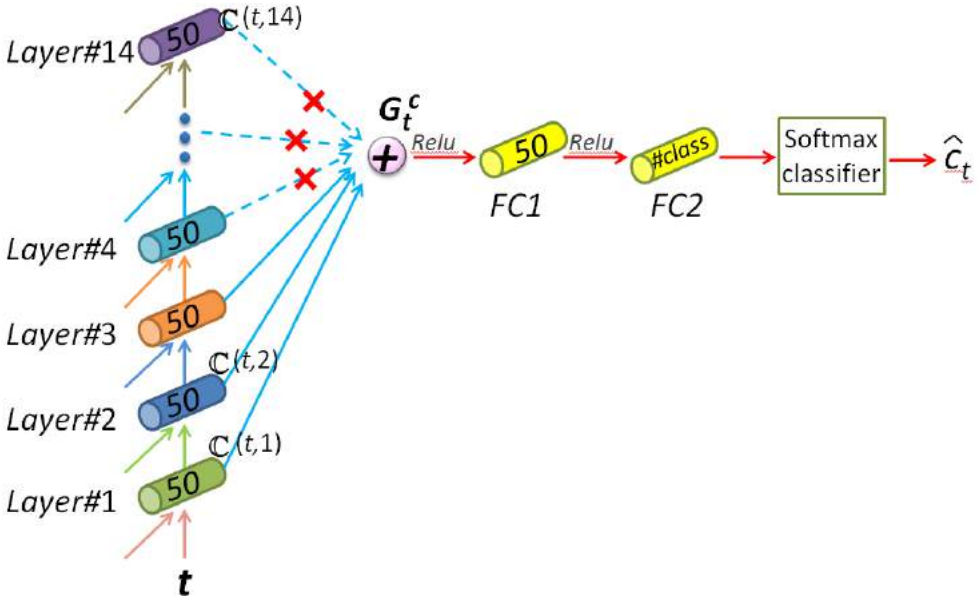
尺度选择网在时间维度上采用一维的卷积来建模不同帧之间的运动动态信息。为了得到一系列不同的时间尺度，该网络采用了扩张卷积（dilated convolution）的设计。通过设计一个层状的（hierarchical）扩张卷积网络的架构，在网络中，不同卷积层的节点拥有不同的感知窗口范围。比如，第 1 个卷积层的感知范围是 2，第 2 层是 4，第 3 层则是 8，等等。

在如上的网络架构中，我们得到了一系列的感知尺度范围。针对在线行为识别不同时间点尺度变化的问题，我们需要在每个时间点来动态选择当前合适的时间窗口尺度。这篇文章设计了一个尺度回归子网络来预测每个时间点需要的尺度大小，该子网络如下图所以。



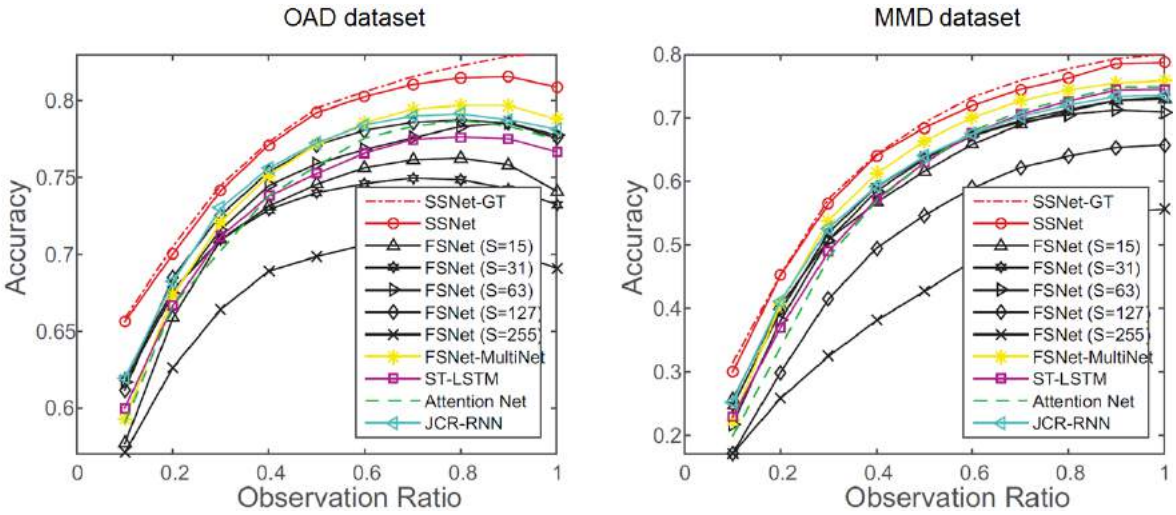
尺度回归子网络通过聚合网络中所有卷积层的信息，并将聚合的信息输入一个全连接网络中，来估计当前动作的当前帧到该动作起始帧之间的距离（s）。得到的 s 则可用于代表当前动作已经执行的部分，因此它可以用于作为预测当前动作类别合适的时间窗口尺度。

得到合适的窗口尺度（s）后，我们可以找到该尺度对应的卷积层。前面我们提到，在尺度选择网中，不同的层对应不同的感知尺度，因此我们找到最匹配的这一层，然后我们就可以使用这层的信息来预测动作类别（c）。这篇论文设计了一个类别预测子网络，在这个子网络中，合适的卷积层的信息被输入全连接网络用于行为预测。如下图所示，假定第 3 卷积层最匹配窗口尺度 s，那么则聚合第 1 到第 3 层的信息。注意论文不仅仅使用了第 3 层信息，还融合了其下面的层，这是因为这种 skip connection 设计可以让网络收敛得更快，同时多尺度的信息融合，也能提高行为预测的准确率。



因为在视频序列的每个时间点上，网络都回归并采用最合适的时间窗口尺度，因此该方法可以得到可靠的预测准确率。值得一提的是，虽然尺度选择网有多个子网络，比如时间序列建模的一维卷积子网络，尺度回归子网络，以及行为预测子网络，但是所有这些子网络均集成在同一个网络架构中，因此整个网络可以端到端进行训练。

作者使用了两个公开数据集来测试尺度选择网的效果，在两个数据集上都取得了很好的实验结果。实验结果如下图所示，其中 SSNet 是本文所提出的尺度选择网；而 SSNet-GT 则表示使用 Ground Truth 尺度来进行行为预测；FS-Net（S）则表示在所有时间点均采用同一个固定的尺度（S）用于行为预测。ST-LSTM 则是本文作者之前发表在 T-PAMI 上的“Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates”。Attention Net 是作者发表在 CVPR17 的“Global Context-Aware Attention LSTM Networks for 3D Action Recognition”。JCR-RNN 是 MSRA 和 PKU 发表在 ECCV16 上的“Online Human Action Detection using Joint Classification-Regression Recurrent Neural Networks”。可以看到，本文提出的 SSNet 的实验结果优于其他方法，并且准确率也接近使用 Ground Truth 尺度的结果。



CVPR-18 阿里巴巴 Spotlight 论文：基于语境对比特征和门控多尺度融合的场景分割

Henghui Ding[1] Xudong Jiang[1] Bing Shuai[1] Ai Qun Liu[1] Gang Wang[2]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

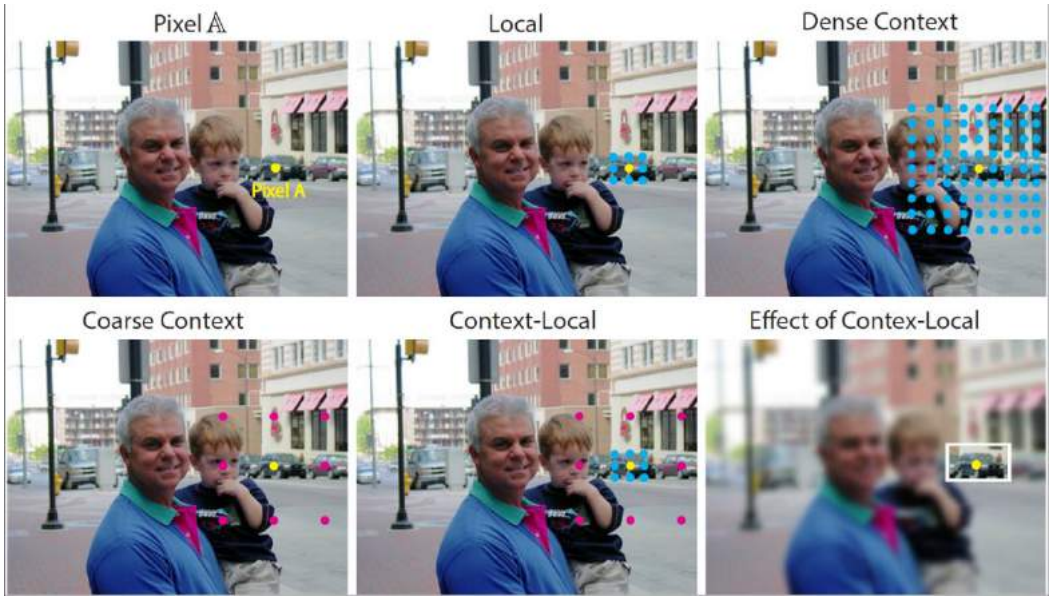
[2]Alibaba AI Labs, Hangzhou, China

本文讨论了场景分割问题，场景分割需要进行像素级别的分类，上下文语境和多尺度特征融合对实现更好的场景分割至关重要。本文首先提出了一种上下文语境和局部信息对比的特征，这种特征不仅利用了信息丰富的上下文语境，而且通过与语境的对比来聚焦更具辨识度的局部信息。这种特征提高了网络的解析性能，尤其提高了对不明显物体和背景填充部分的分割效果。同时，本文提出了一种门控融合机制，不同于以往的多尺度特征融合，门控融合可以根据输入图像的特征表象来为不同位置的分类选择性地融合多尺度特征。门控的值由本文提出的网络产生，会随输入图像的变

化而变化。这种门口融合机制可以控制不同尺度特征的信息流动，使网络对不同尺度的物体有更强的适应力。本文提出的模型在 Pascal Context, SUN-RGBD 和 COCO Stuff 三个场景分割数据集上验证了性能，取得了目前最高的场景分割性能。

本文致力于场景分割中有两大问题：场景图片中像素形式的多样化（例如，显著或者不显著，前景或者背景），场景图片中物体大小的多样性。并针对这两个问题分别提出了语境对比局部特征和门控多尺度融合。

1. 语境对比局部特征



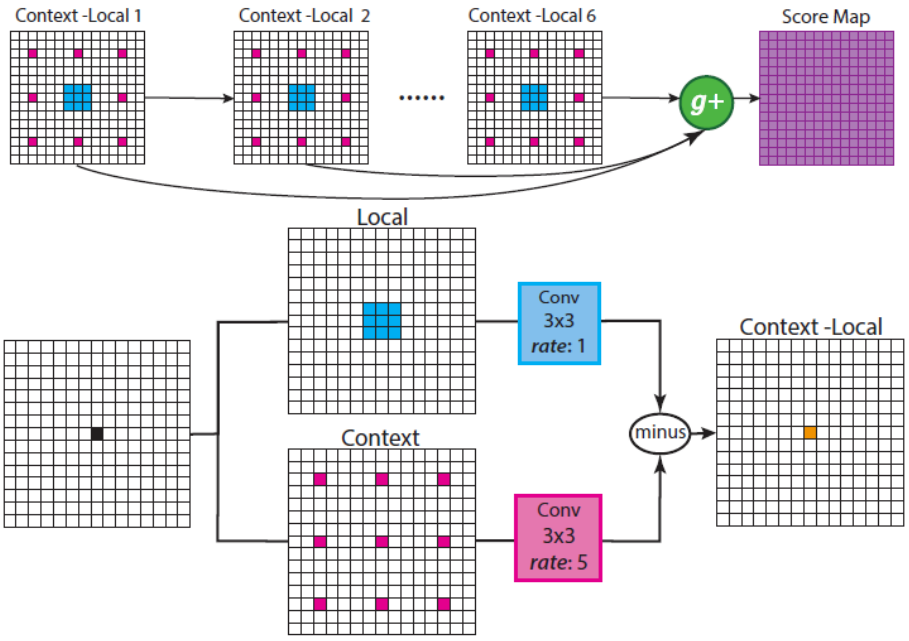
图一

场景图片中物体形式具有多样化，如显著或者不显著，前景或者背景。图像分类问题一般仅关注于图像中最显著的物体，而场景分割需要对所有像素进行分类，所以在进行场景分割时应该对不同存在形式的物体都给予关注，尤其是不显著的物体和背景。

上下文语境信息对于实现良好的场景分割至关重要。然而语境信息容易被场景图片中的显著物体的特征所主导，导致场景中其他的不显著物体和背景的特征被弱化甚至忽略。如图一所示，像素 A 属于不显著物体。像素 A 的局部特征（Local）包含了像素 A 的主要信息，但是缺乏足够的全局信息（如 路），不能实现稳定的分割。但是收集语境信息（Context），就会带来旁边显著物体（人）的特征信息，导致像素 A 的语境特征被人的特征信息所主导，像素 A 自身的特征信息被弱化甚至忽略。

为此，本文提出了语境对比局部特征，同时收集像素 A 的局部特征和全局语境特征（如图二所示），然后将两者进行对比融合，一方面可以保护并突出局部特征，另一方面充分利用了信息丰富的语境特征。语境对比的局部特征，包含了富含信息的全局特征以及对像素 A 至关重要的局部特征，

并且使得像素 A 的特征不会被其他显著物体特征所主导。语境对比局部特征的效果如图一中最后一幅图片所示。同时，在最终模型 CCL 中，多个 context-local 被连接起来以获得多层次的语境对比局部特征，如图二所示。

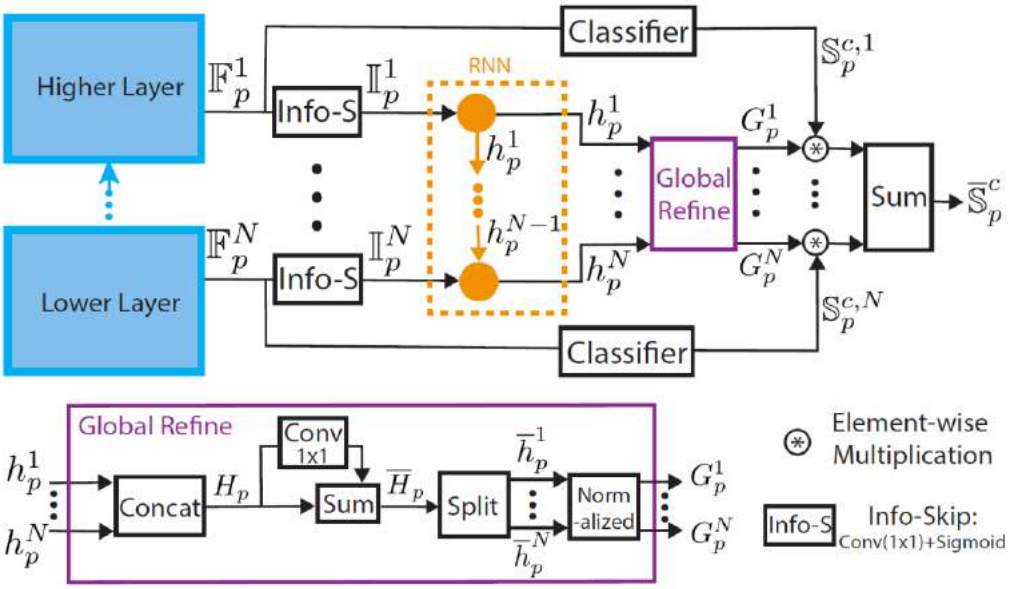


图二

2. 门控多尺度融合

场景分割的数据集中存在着大量的不同大小的物体，不同大小的物体所需的感受野和特征尺度不同，用单一尺度的特征对所有像素进行分类是不合理的，因此需要进行多尺度特征融合。本文采用了 FCN 网络中的 skip 结构来获取 DCNN 不同特征层的特征信息，但 FCN 中对 skip 的结果简单相加融合，并没有考虑不同 skip 结果的差异性。不同于 FCN 的是，本文提出了一种门控机制来进行多尺度特征的选择性融合。门控多尺度融合根据特征尺度、语境等信息来动态决定图像中每个像素最适合的感受野，对 skip 的分割结果进行选择性融合。门控的值由本文提出的网络产生（如图三所示），网络根据输入图像的特征表象生成相应的门控值，由这些门控的值来决定不同层的 skip 以多大比例进行融合，控制网络信息流动，产生最终的预测。

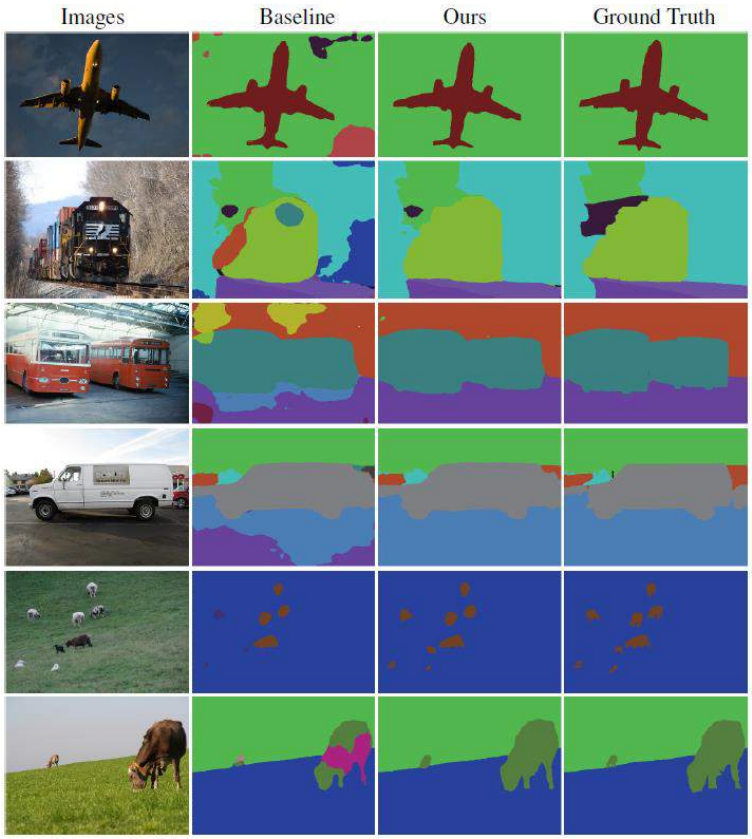
通过门控多尺度融合，网络可以为每个像素选择一个合理的组合方案，将合适的特征增强并将不合适的特征进行抑制。在门控多尺度融合中，可以添加更多的 skip 来提取更丰富的多尺度特征信息，同时不用担心 skip 中有不好的结果。这种门口融合机制可以控制不同尺度特征的信息流动，使网络对不同尺度的物体有更强的适应力。



图三

3. 实验

场景分割效果对比如图四所示，本文提出的方法对不显著物体和背景的分割效果有明显提升，同时对多尺度物体有很强的适应力。



图四

CVPR-18 阿里巴巴 Spotlight 论文：所见所想所找－基于生成模型的跨模态检索

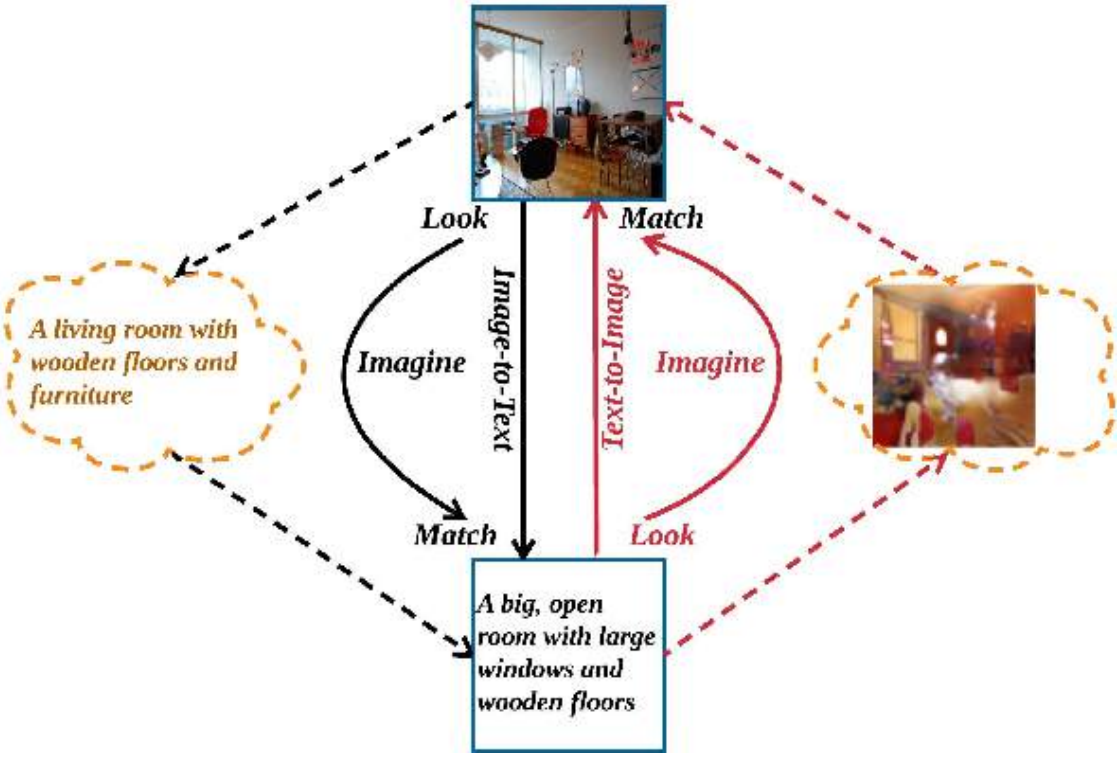
1. 摘要

视觉-文本跨模态检索已经成为计算机视觉和自然语言处理领域结合的一个热点。**对于跨模态检索而言，如何学到合适的特征表达非常关键**。本文提出了一种基于生成模型的跨模态检索方法，该方法可以学习跨模态数据的高层次特征相似性，以及目标模态上的局部相似性。本文通过大量的实验证明了所提出的方法可以准确地匹配图像和文本，并且在 MSCOCO 以及 Flickr30K 的数据集上都取得了 state-of-the-art 的效果。

2. 引言

我们已经进入到了一个大数据时代，不同模态的数据例如文本、图像等正在以爆炸性的速度增长。这些异质的模态数据也给用户的搜索带来了挑战。对于文本-视觉的跨模态表示，常见的方法就是首先每个模态的数据编码成各自模态的特征表示，再映射到一个共同空间内。通过 ranking loss 来对其进行优化，使得相似的图像-文本对映射出的特征向量之间的距离小于不相似的图像-文本对之间的距离。尽管这种方法学习出的特征可以很好地描述多模态数据高层语义，但是没有充分地挖掘图像的局部相似度和句子的句子层次相似度。例如文本检索图片时，我们会更多地关注图片的颜色、纹理以及布局等细节信息。而仅仅进行高层次特征匹配，显然无法考虑到局部的相似度。

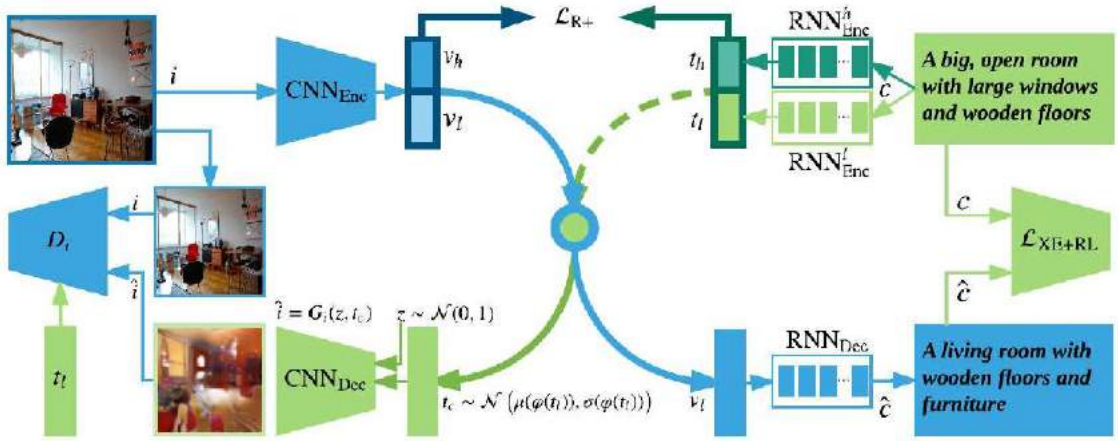
本文的想法来源于对人的思维的思考。对于人来说，给定一段文字描述去检索匹配的图像，一名训练有素画家可以比普通人找到更匹配的图像，那是因为画家知道预期的图片是什么样；类似，给一幅图片去检索匹配的文字描述，一名作家也往往会给出比普通人更好的描述。我们把这种对检索目标有预期的过程称为--”Imagine”或者“脑补”。因此，我们提出了一种基于生成模型的跨模态特征学习框架（generative cross-modal feature learning framework，GXN），下图展示了本文的思想：



我们把原来的 Look 和 Match 变成了三个步骤：Look, Imagine 和 Match，也称为”所看所想所找”。Look 叫“所看”，“看”是理解，实际就是提取特征。Imagine 叫“所想”，根据“所看”去“脑补”预期的匹配结果，也就是从得到的局部特征去生成目标模态的数据；Match 也叫“所找”，根据生成/脑补的结果进行局部层次（sentence-level/pixel-level）匹配和高层次语义特征匹配。

3. 方法

GXN 包括三个模块：多模态特征表示（上部区域）；图像-文本生成特征学习（蓝色路径）和文本-图像生成对抗特征学习（绿色路径）。



- 第一个部分（上部区域）和基本的跨模态特征表示做法类似，将不同模态的数据映射到共同空间。这里包括一个图像编码器CNN_{Enc}和两个句子编码器$\text{RNN}^h_{\text{Enc}}$和$\text{RNN}^l_{\text{Enc}}$。之所以分开 2 个句子编码器，是便于学到不同层次的特征。其中，(v_h, t_h)是高层语义特征而(v_l, t_l)作为局部层次的特征。这里的局部层次特征是通过生成模型学习得到的。

- 第二部分（蓝色路径）从底层视觉特征v_l生成一个文本描述。包括一个图像编码器CNN_{Enc}和一个句子解码器RNN_{Dec}。这里计算损失时我们结合了增强学习的思想，通过奖励的方式来确保生成句子和真实句子之间具有最大的相似度。

- 第三部分（绿色路径）通过使用一个 cGAN 从文本特征t_l中生成一幅图像，包括一个生成器CNN_{Dec}和一个判别器D_i。判别器用来区分基于文本生成的图像与真实图像。

最终，我们通过两路的跨模态特征生成学习学习到更好的跨模态特征表示。在测试时，我们只需要计算$\{v_h, t_h\}$和$\{v_l, t_l\}$之间的相似度来进行跨模态检索。

4. 实验

本文提出的方法在 MSCOCO 数据集上和目前前沿的方法进行比较，并取得了 state-of-the-art 的结果。

Model	Image-to-Text Retrieval			Text-to-Image Retrieval			Sum
	R@1	R@10	Med r	R@1	R@10	Med r	
	1K Test Images						
Skip-thought vectors [15]	33.8	82.1	3.0	25.9	74.6	4.0	216.4
DVSA [11]	38.4	80.5	**1.0**	27.4	74.8	3.0	221.1
Fisher Vector [16]	39.4	80.9	2.0	25.1	76.6	4.0	222.0
m-RNN [21]	41.0	83.5	2.0	29.0	77.0	3.0	230.5
RNN+Fisher Vector [18]	40.8	83.2	2.0	29.6	82.8	3.0	236.4
MNLM [14]	43.4	85.8	2.0	31.0	79.9	3.0	240.1
m-CNN [20]	42.8	84.1	2.0	32.6	82.8	3.0	242.3
HM-LSTM [26]	43.9	87.8	2.0	36.1	86.7	3.0	254.5
Order-embeddings [38]	46.7	88.9	2.0	38.9	85.9	2.0	260.4
DSPE+Fisher Vector [39]	50.1	89.2	-	39.6	86.9	-	265.8
sm-LSTM [9]	53.2	91.5	**1.0**	40.7	87.4	2.0	272.8
*VSE++ (ResNet152, fine-tune) [3]	64.7	95.9	**1.0**	52.0	92.0	**1.0**	304.6
Gen-XRN (i2t+t2i)	**68.5**	**97.9**	**1.0**	**56.6**	**94.5**	**1.0**	**317.5**
	5K Test Images						
Order-embeddings [38]	23.3	65.0	5.0	18.0	57.6	7.0	163.9
*VSE++ (ResNet152, fine-tune) [3]	41.3	81.2	**2.0**	30.3	72.4	4.0	225.2
Gen-XRN(t2i+t2i)	**42.0**	**84.7**	**2.0**	**31.7**	**74.6**	**3.0**	**233.0**

5. 总结

本文创新性地将图像-文本生成模型和文本-图像生成模型引入到传统的跨模态表示中，使其不仅能学习到多模态数据的高层的抽象表示，还能学习到底层的表示。显著超越 state-of-the-art 方法的表现证实了该方法的有效性。

CVPR2018 阿里巴巴 Poster 论文：整体还是局部？应用 Localized GAN 进行图像内容编辑、半监督训练和解决 mode collapse 问题

1. 摘要

GAN 自诞生以来吸引了众多相关的研究，并在理论、算法和应用方面取得了很多重大的突破。我们试图从一个全新的几何角度，用**局部的**观点建立一种与之前经典 GAN 模型所采用的**整体方法**不同的理论和模型，并以此建立和半监督机器学习中 Laplace-Beltrami 算子的联系，使之不再局限于传统的图模型(Graph)方法，并在用少量标注样本训练深度学习模型上取得了优异的性能；同时我们还展示了如果用 Localized GAN (LGAN)对给定图像在局部坐标系下进行编辑修改，从而获得具有不同角度、姿态和风格的新图像；我们还将进一步揭示如何从流型切向量独立性的角度来解释和解决 GAN 的 mode collapse 问题。

该工作由 UCF 齐国君教授领导的 UCF **MAPLE** 实验室(**MA**chine Perception and **LE**arning)和阿里巴巴华先胜博士领导的城市大脑机器视觉研究组合作完成，并将发表在 CVPR 2018 上。

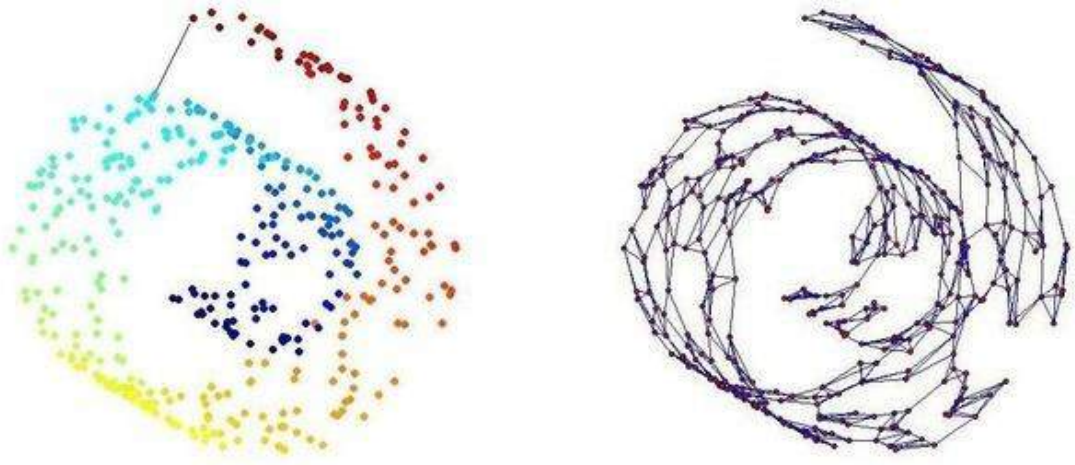
Global versus Localized Generative Adversarial Nets

Guo-Jun Qi, Liheng Zhang, Hao Hu, Marzieh Edraki
guojun.qi@ucf.edu, {lihengzhang1993,hao_hu,m.edraki}@knights.ucf.edu
Jingdong Wang, and Xian-Sheng Hua
welleast@outlook.com, xiansheng.hxs@alibaba-inc.com
Laboratory for MAchine Perception and LEarning (MAPLE)
maple.cs.ucf.edu
University of Central Florida and Alibaba Group

Guo-Jun Qi, Liheng Zhang, Hao Hu, Marzieh Edraki, Jingdong Wang and Xian-Sheng Hua. Global versus Localized Generative Adversarial Nets, in CVPR 2018.

2. GAN 和基于图模型的半监督机器学习的关系

GAN 除了用来生成数据，我们认为一个非常重要作用是：我们第一次有了一个比较理想的工具，可以用来表示和描述数据流型(manifold)。之前，如果我们想表示流型，一般是借助于一个图模型（Graph）。在图模型里，我们用节点表示数据点，用边表示数据直接的相似性。有了 Graph，我们可以定量计算数据点上函数的变化。比如，在分类问题中，我们感兴趣的函数是分类函数，输出的是数据点的标签。有了基于 Graph 的流型，我们就可以建立一个分类模型：**它输出的分类标签在相似样本上具有最小的变化**。这个就是一种平滑性的假设，是基于图的半监督方法的核心假设。



上图：基于图的流型表示和半监督分类。

尽管这种基于图的半监督方法取得了很大的成功，但是它的缺点也是很明显的。当数据点数量非常巨大的时候，构建这样一个 Graph 的代价会非常大。为了解决这个问题， Graph 为我们提供了一个很好的基础。通过训练得到的生成器 G(z)，其实就是一个非常好的流型模型。这里 z 就是流型上的参数坐标，通过不断变化 z，我们就可以在高维空间中划出一个流型结构。

有了这样一个流型和它的描述 G，我们可以在数据流型上研究各种几何结构。比如切向量空间、曲率，进而去定义**在流型上，沿着各个切向量，函数会如何变化**等等。好了，这里 GAN 就和半监督学习联系起来了。以前我们是用 Graph 这种离散的结果去研究分类函数的变化，并通过最小化这种变化去得到平滑性假设。

现在，有了流型直接的参数化描述 G(z)，我们就能直接去刻画一个函数（比如分类问题中的分类器）在流型上的变化，进而去建立一个基于这种参数化流型的半监督分类理论，而非去借助基于图的流型模型。

具体来说，半监督图流型中，我们常用到 Laplacian 矩阵来做训练；现在，有了参数化的流型后，我们就可以直接定义 Laplace-Beltrami 算子，从而实现半监督的训练。下面是基于这个方法在一些数据集上得到的结果。更多的结果可以参考我们的论文“Global versus Localized Generative Adversarial Networks“。

Method	SVHN	CIFAR-10
Π model [12]	4.82 ± 0.17	12.36 ± 0.31
Temporal Ensembling [12]	4.42 ± 0.16	12.16 ± 0.24
Sajjadi et al. [7]	-	11.29 ± 0.24
VAT [12]	**3.86**	10.55
VadD [6]	4.16 ± 0.08	11.32 ± 0.11
Our approach	4.39 ± 0.14	$\mathbf{9.77 \pm 0.13}$

Table 3. Classification errors on both SVHN and CIFAR-10 with 1000 and 4000 labeled training examples respectively. The best result is highlighted in bold.

Method	CIFAR-10	CIFAR-100
Π model [12]	12.36 ± 0.31	39.19 ± 0.36
Temporal Ensembling [12]	12.16 ± 0.24	38.65 ± 0.51
Sajjadi et al. [22]	11.29 ± 0.24	-
VAT [14]	10.55	-
VadD [16]	11.32 ± 0.11	-
LGAN (Conv-Large)	$\mathbf{9.77 \pm 0.13}$	$\mathbf{35.52 \pm 0.33}$

Table 4. Classification errors on both CIFAR-10 and CIFAR-100 with 4,000 and 10,000 labeled training examples respectively. The best result is highlighted in bold.

上表：在 SVHN, CIFAR-10 和 CIFAR-100 上的半监督学习效果。

3. 用全局还是局部坐标来研究 GAN

这里，有个比较精细的问题。通常的 GAN 模型，得到的是一个全局的参数话模型：我们只有一个 z 变量去参数化整个流型。事实上，在数学上，这种整体的参数化王是不存在的，比如我们无法用一个参数坐标去覆盖整个球面。这时我们往往要借助于通过若干个局部的坐标系去覆盖整个流型。

同时，使用局部坐标系的另一个更加实际的好处是，我们给定一个目标数据点 x 后，整体坐标系 G(z)要求我们必须知道对应的一个参数坐标 z；而使用局部坐标系后，我们就直接可以在 x 附近去建立一个局部坐标系 G(x, z)去研究流型周围的几何结构，而不用去解一个逆问题去去它对应的 z 了。这个极大地方便了我们处理流型上不同数据点。

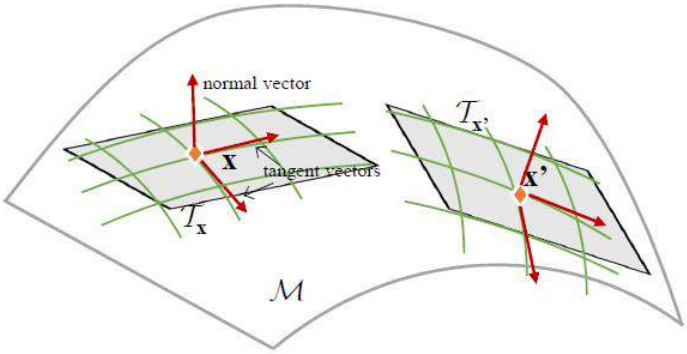


Figure 1. Illustration of a curved manifold $\mathcal{M}$ embedded in 3-dimensional ambient space. At each location $\mathbf{x}$, its tangent space $\mathcal{T}_\mathbf{x}$ consists of all tangent vectors to the manifold. These tangent vectors characterize the geometry of local transformations allowed to move a point $\mathbf{x}$ on $\mathcal{M}$.

图：流型的局部参数化表示。

沿着这个思路，我们可以利用参数化的局部坐标和它表示的流型来研究一系列问题。

1. 比较理论的研究可以专注于，有了这些局部参数表示，如何去定义出一整套黎曼流型的数学结构，比如局部的曲率，黎曼度量，和如果沿着流型去算测地线和两个数据点之间的测地距离。

2. 从应用的角度，给定了一个图像 x，用局部表示 G(x, z)可以对这个 x 在它的局部领域中做各种编辑操作或者控制图像的各种属性，从而可以帮助我们生成想要的图像；比如不同角

度的人脸、人体姿态、物体，甚至不同风格、表现不同情感的图像等等。这在安防、内容生成、虚拟现实等领域都会有广泛的应用前景。

4. 从几何角度研究 Mode collapse 问题

当然，从几何和流型参数化的角度还可以给出对 GAN 更深入的理解，比如对 mode collapse 问题。在 GAN 的相关研究中，mode collapse 是一个被广泛关注的问题。有很多相关的论文在从不同角度来研究和解决这个问题。

而基于 Localized GAN 所揭示的几何方法，我们可以从流型局部崩溃的角度来

解释和避免 **GAN 的 mode collapse**。具体来说，给定了一个 z，当 z 发生变化的时候，对应的 G(z)没有变化，那么在这个局部，GAN 就发生了 mode collapse，也就是不能产生不断连续变化的样本。这个现象从几何上来看，就是对应的流型在这个局部点处，沿着不同的切向量方向不再有变化。换言之，所有切向量不再彼此相互独立--某些切向量要么消失，要么相互之间变得线性相关，从而导致流型的维度在局部出现缺陷（dimension deficient）。

为了解决这个问题，最直接的是我们可以给流型的切向量加上一个正交约束(Orthonormal constraint)，从而避免这种局部的维度缺陷。下图是在 CelebA 数据集上得到的结果。可以看到，通过对不同的切向量加上正交化的约束，我们可以在不同参数方向上成功地得到不同的变化。

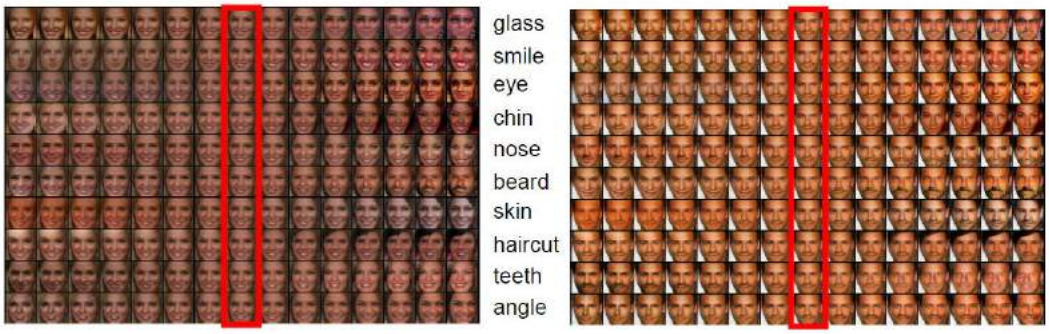


Figure 2. Faces generated by LGAN on the CelebA dataset. The middle column in a red bounding box represents the image at the origin $\mathbf{z} = 0$ of a local coordinate chart. In each row, the images are generated along a local coordinate. There exist various patterns of image variations across different rows of faces, including whether wearing glasses and the variations in expressions, eyes, haircuts and so forth.

上图：在给定输入图像的局部坐标系下对人脸的不同属性进行编辑。

值得注意的是，尽管我们是从局部 GAN 的角度推导和实现了对切向量的正交化约束，**这个思路和方法同样适用于传统的整体 GAN 模型**。我们只需要在训练整体 GAN 模型的同时，在每个训练数据样本或者一个 batch 的子集上也加上这个约束来求取相应的下降梯度就同样可以训练整体 GAN 模型；这个方向可以引申出未来的相关工作。

论文原文地址：https://arxiv.org/pdf/1711.06020.pdf

CVPR2018 阿里巴巴 Poster 论文：处理多种退化类型的卷积超分辨率

Learning a Single Convolutional Super-Resolution Network for Multiple Degradations

Kai Zhang[1,2,3], Wangmeng Zuo[1], Lei Zhang[2]
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[3]DAMO Academy, Alibaba Group
cskaizhang@gmail.com, wmzuo@hit.edu.cn, cslzhang@comp.polyu.edu.hk

1. 摘要

近年来，深度卷积神经网络（CNN）方法在单幅图像超分辨率（SISR）领域取得了非常大的进展。然而现有基于 CNN 的 SISR 方法主要假设低分辨率（LR）图像由高分辨率（HR）图像经过双三次(bicubic)降采样得到，因此当真实图像的退化过程不遵循该假设时，其超分辨结果会非常差。此外，现有的方法不能扩展到用单一模型解决多种不同的图像退化类型。为此，提出了一种维度拉伸策略使得单个卷积超分辨率网络能够将 SISR 退化过程的两个关键因素（即模糊核和噪声水平）作为网络输入。归因于此，训练得到超分辨网络模型可以处理多个甚至是退化空间不均匀的退化类型。实验结果表明提出的卷积超分辨率网络可以快速、有效的处理多种图像退化类型，为 SISR 实际应用提供了一种高效、可扩展的解决方案。

2. 引言

单幅图像超分辨率（SISR）的目的是根据单幅低分辨（LR）图像输入得到清晰的高分辨率（HR）图像。一般来说，LR 图像$\boldsymbol{y}$是清晰 HR 图像$\boldsymbol{x}$由下面的退化过程得来，

$$\boldsymbol{y} = (\boldsymbol{x} \otimes \boldsymbol{k}) \downarrow_s + \boldsymbol{n}$$

其中$\boldsymbol{x} \otimes \boldsymbol{k}$表示 HR 清晰图像$\boldsymbol{x}$与模糊核$\boldsymbol{k}$之间的卷积，$\downarrow_s$表示系数为$s$的降采样算子，$\boldsymbol{n}$表示标准差（噪声水平）为 σ 的加性高斯白噪声（AWGN）。

SISR 方法主要分为三类：基于插值的方法、基于模型的方法以及基于判别学习的方法。基于插值的方法（例如：最近邻插值、双三次插值）虽然速度快，但是其效果比较差。基于模型的方法通过引入图像先验，例如：非局部相似性先验、去噪先验等，然后求解目标函数得到视觉质量较好的 HR 图像，然而速度较慢。虽然结合基于 CNN 的去噪先验可以在某种程度上提升速度，但仍然受限于一些弊端，例如：无法进行端对端的训练，包含一些比较难调的参数等。基于判别学习

的方法尤其是基于 CNN 的方法因其速度快、可以端对端的学习因而效果好等在近几年受到了广泛关注，并且逐渐成为解决 SISR 的主流方法。

自从首个用 CNN 解决 SISR 的工作 SRCNN 在 ECCV（2014）发表以来，各种不同的改进方法相继提出。例如，VDSR 在 PSNR 指标上取得了非常大的提升；ESPCN 和 FSRCNN 分别在速度上进行了改进；SRGAN 在放大倍数较大情况下针对视觉效果的改善提出了有效的方法。然而这些方法都存在一个共同缺点，也就是它们只考虑双三次(bicubic)降采样退化模型并且不能灵活的将其模型扩展到同时（非盲）处理其他退化类型。由于真实图像的退化过程多种多样，因而此类方法的有效实际应用场景非常有限。一些 SISR 工作已经指出图像退化过程中的模糊核的准确性对 SISR 起着至关重要的作用，然而并没有基于 CNN 的相关工作将模糊核等因素考虑在内。为此引出本文主要解决的问题：是否可以设计一个非盲超分辨率（non-blind SISR）模型用以解决不同的图像退化类型？

3. 方法

本文首先分析了在最大后验（MAP）框架下的 SISR 方法，借此希望可以指导 CNN 网络结构的设计。由于 SISR 问题的不适定性，通常需要引入正则项来约束解空间。具体来说，LR 图像$\boldsymbol{y}$对应的 HR 图像$\boldsymbol{x}$可以通过求解下述问题近似，

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \frac{1}{2\sigma^2}\|(\boldsymbol{x} \otimes \boldsymbol{k})\downarrow_s - \boldsymbol{y}\|^2 + \lambda\Phi(\boldsymbol{x})$$

其中$\frac{1}{2\sigma^2}\|(\boldsymbol{x}\otimes\boldsymbol{k})\downarrow_s - \boldsymbol{y}\|^2$为似然（也即数据保真）项，$\Phi(\boldsymbol{x})$为先验（也即正则）项，$\lambda$为似然项和先验项之间的权衡参数。简单来说，上述公式包含两点：1）估计得到的 HR 图像不仅要符合 SISR 的退化过程，并且还要满足清晰图像所具有的先验特征；2）对于非盲超分辨率问题，$\boldsymbol{x}$的求解与 LR 图像$\boldsymbol{y}$、模糊核$\boldsymbol{k}$、噪声水平σ以及权衡参数 λ 有关。简而言之，非盲 SISR 的 MAP 估计可以表示为 $\hat{\boldsymbol{x}} = \mathcal{F}(\boldsymbol{y}, \boldsymbol{k}, \sigma, \lambda; \Theta)$，其中$\Theta$为 MAP 估计中的参数。进而如果将 CNN 看作 MAP 估计另一种形式的解，那么有如下结论：

1) 由于数据保真项对应着 SISR 的退化过程，因此退化过程的准确建模对 SISR 的结果起着至关重要的作用。然而现有的基于 CNN 的方法其目标是求解下面的问题，

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\downarrow_s - \boldsymbol{y}\|^2 + \Phi(\boldsymbol{x})$$

由于没有将模糊核和噪声等因素考虑在内，因此其实用性非常有限。

2) 为了设计更加有效的基于 CNN 的 SISR 模型，应该将更多的图像退化类型考虑在内，一个简单的思路就是将模糊核$\boldsymbol{k}$和噪声水平σ也作为网络的输入。由于权衡参数λ可以融入噪声水平σ之中，因此 CNN 映射函数可以简化成如下形式：

$$\hat{\boldsymbol{x}} = \mathcal{F}(\boldsymbol{y}, \boldsymbol{k}, \sigma; \Theta)$$

3) 由于 MAP 估计中大部分的参数都对应着图像先验部分，而图像先验是与图像退化过程不相关的，因此单一的 CNN 模型具有处理不同退化类型的建模能力。

通过上述分析可以得出非盲 SISR 应该将退化模型中的模糊核和噪声水平也作为网络的输入。然而 LR 图像、模糊核和噪声水平三者的维度是不同的，因此不能直接作为 CNN 的输入。为此本文提出了一种维度拉伸策略。假设 LR 图像大小为$\mathrm{W}\times\mathrm{H}$，首先将向量化的模糊核 PCA 降维，然后和噪声水平并在一起得到一个$\mathrm{t}+1$维的向量$\boldsymbol{v}$，接着将$\boldsymbol{v}$拉伸为$\mathrm{W}\times\mathrm{H}\times(\mathrm{t}+1)$维的张量，我们将此张量称之为退化图（Degradation Maps），其中第i个$\mathrm{W}\times\mathrm{H}$图的所有元素均为$\boldsymbol{v}_i$。

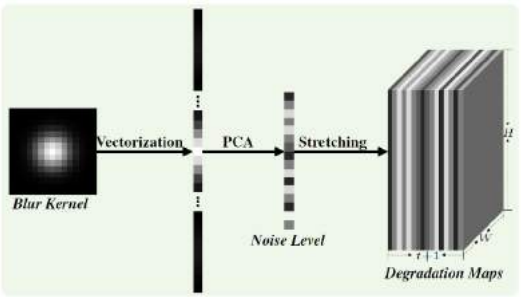


图 1：维度拉伸示意图。

至此，我们可以将退化图和 LR 图像合并在一起作为 CNN 的输入。为了证明此策略的有效性，选取了快速有效的 ESPCN 超分辨网络结构框架。值得注意的是为了加速训练过程的收敛速度，同时考虑到 LR 图像中包含高斯噪声，因此网络中加入了 Batch Normalization 层。

图 2 给出了提出的超分辨率网络（简称 SRMD）结构框架。

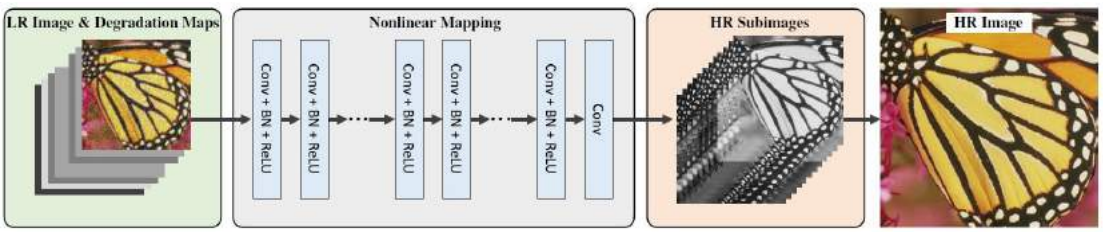


图 2：提出的超分辨率网络结构框架（卷积层数为 12，每层通道数为 128）。

4. 实验

在训练阶段，SRMD 采用了各向同性和各向异性的高斯模糊核、噪声水平在[0, 75]之间的高斯白噪声以及 bicubic 降采样算子。需要指出的是 SRMD 可以扩展到其他降采样算子，甚至其它退化模型。

在测试阶段，SRMD 比较了不同方法在同为 bicubic 降采样退化下的 PSNR 和 SSIM 结果（如表 1 所示）。可以看出虽然 SRMD 是用来处理各种不同的退化类型，但是仍然在 bicubic 降采样退化下取得不错的效果。需要指出的是 SRMD 在速度上也有很大的优势，在 Titan Xp GPU 上处理 512×512 的 LR 图像仅需 0.084 秒，是 VDSR 超分辨率两倍所用时间的一半。表 2 给出了不同退化类型下的 PSNR 和 SSIM 结果比较，可以看到 SRMD 同样取得了不错的效果。图 4 举例说明了 SRMD 可以设定非均匀退化图，进而可以处理退化空间不均匀的 LR 图像。最后，图 5 展示了不同方法在真实图像上的视觉效果比较，可以看到 SRMD 复原的 HR 图像在视觉效果上明显优于其它方法。

Dataset	Scale Factor	Bicubic	SRCNN [9]	VDSR [24]	SRResNet [29]	DRRN [44]	LapSRN [27]	SRMD	SRMDNF
		PSNR / SSIM							
Set5	×2	33.64 / 0.929	36.62 / 0.953	37.56 / 0.959	–	37.66 / 0.959	37.52 / 0.959	37.53 / 0.959	37.79 / 0.960
	×3	30.39 / 0.868	32.74 / 0.908	33.67 / 0.922	–	33.93 / 0.923	33.82 / 0.922	33.86 / 0.923	34.12 / 0.925
	×4	28.42 / 0.810	30.48 / 0.863	31.35 / 0.885	32.05 / 0.891	31.58 / 0.886	31.54 / 0.885	31.59 / 0.887	31.96 / 0.893
Set14	×2	30.22 / 0.868	32.42 / 0.906	33.02 / 0.913	–	33.19 / 0.913	33.08 / 0.913	33.12 / 0.914	33.32 / 0.915
	×3	27.53 / 0.774	29.27 / 0.821	29.77 / 0.832	–	29.94 / 0.834	29.89 / 0.834	29.84 / 0.833	30.04 / 0.837
	×4	25.99 / 0.702	27.48 / 0.751	27.99 / 0.766	28.49 / 0.780	28.18 / 0.770	28.19 / 0.772	28.15 / 0.772	28.35 / 0.777
BSD100	×2	29.55 / 0.843	31.34 / 0.887	31.89 / 0.896	–	32.01 / 0.897	31.80 / 0.895	31.90 / 0.896	32.05 / 0.898
	×3	27.20 / 0.738	28.40 / 0.786	28.82 / 0.798	–	28.91 / 0.799	28.82 / 0.798	28.87 / 0.799	28.97 / 0.803
	×4	25.96 / 0.667	26.90 / 0.710	27.28 / 0.726	27.58 / 0.735	27.35 / 0.726	27.32 / 0.727	27.34 / 0.728	27.49 / 0.734
Urban100	×2	26.66 / 0.841	29.53 / 0.897	30.76 / 0.914	–	31.02 / 0.916	30.82 / 0.915	30.89 / 0.916	31.33 / 0.920
	×3	24.46 / 0.737	26.25 / 0.801	27.13 / 0.828	–	27.38 / 0.833	27.07 / 0.828	27.27 / 0.833	27.57 / 0.840
	×4	23.14 / 0.657	24.52 / 0.722	25.17 / 0.753	–	25.35 / 0.758	25.21 / 0.756	25.34 / 0.761	25.68 / 0.773

表 1：不同方法在 bicubic 降采样退化下的 PSNR 和 SSIM 结果比较（其中 SRMDNF 表示不考虑噪声情况下训练得到的模型）。

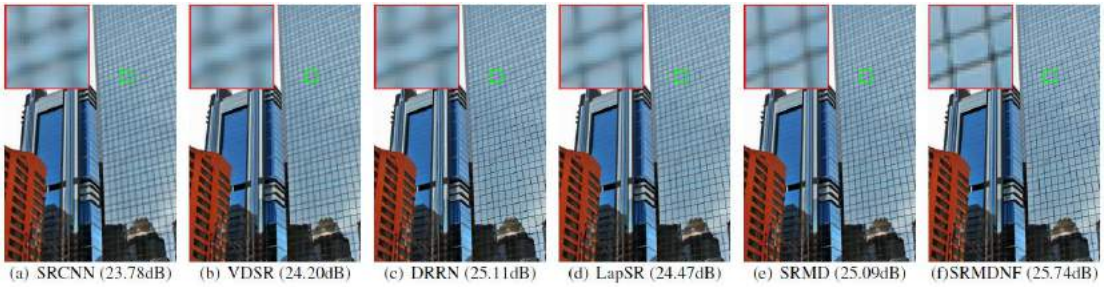


图 3：不同方法在 bicubic 降采样退化下超分辨率四倍的视觉效果比较。

Degradation Settings			VDSR [24]	NCSR [11]	IRCNN [57]	DnCNN [56]+SRMDNF	SRMD	SRMDNF
Kernel Width	Down-sampler	Noise Level	PSNR (×2/×3/×4)					
0.2	Bicubic	0	37.56/33.67/31.35	– /23.82/–	37.43/33.39/31.02	–	37.53/33.86/31.59	37.79/34.12/31.96
0.2	Bicubic	15	26.02/25.40/24.70	–	32.60/30.08/28.35	32.47/30.07/28.31	32.76/30.43/28.79	–
0.2	Bicubic	50	16.02/15.72/15.46	–	28.20/26.25/24.95	28.20/26.27/24.93	28.51/26.48/25.18	–
1.3	Bicubic	0	30.57/30.24/29.72	– /21.81/–	36.01/33.33/31.01	–	37.04/33.77/31.56	37.45/34.16/31.99
1.3	Bicubic	15	24.82/24.70/24.30	–	29.96/28.68/27.71	27.68/28.78/27.71	30.98/29.43/28.21	–
1.3	Bicubic	50	15.89/15.68/15.43	–	26.69/25.20/24.42	24.35/25.19/24.39	27.43/25.82/24.77	–
2.6	Bicubic	0	26.37/26.31/26.28	– /21.46/–	32.07/31.09/30.06	–	33.24/32.59/31.20	34.12/33.02/31.77
2.6	Bicubic	15	23.09/23.07/22.98	–	26.44/25.67/24.36	– /21.33/23.85	28.48/27.55/26.82	–
2.6	Bicubic	50	15.58/15.43/15.23	–	22.98/22.16/21.43	– /19.03/21.15	25.85/24.75/23.98	–
1.6	Direct	0	– /30.54/ –	– /33.02/ –	– /33.38/ –	–	– /33.74/ –	– /34.01/ –

表 2：不同方法在不同退化类型下的 PSNR 和 SSIM 结果比较。

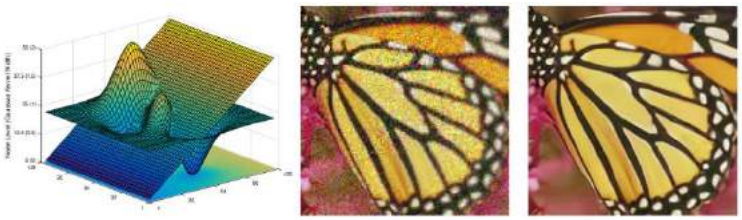


图 4：举例说明 SRMD 可以处理退化空间不均匀的情形。（a）噪声水平以及模糊核宽度的空间分布；（b）LR 图像（最近邻插值放大）；（c）复原得到的 HR 图像（放大两倍）。

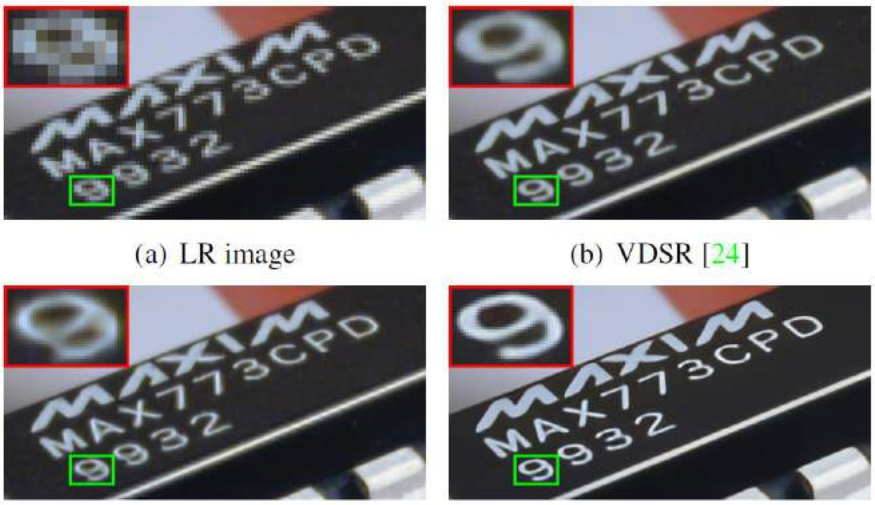


图 5：不同方法在 SISR 经典测试图像“Chip”上超分辨率四倍的视觉效果比较。

5. 结论

最后总结一下，本文的主要贡献有三个方面：

- 提出了一种简单、有效、可扩展的超分辨率模型，其不仅可以处理 bicubic 降采样退化模型，并且可以处理多个甚至是退化空间不均匀的退化类型，为 SISR 实际应用提供了一种解决方案。
- 提出了一种简单有效的维度拉伸策略使得卷积神经网络可以处理维度不同的输入，此策略可以扩展到其他应用。
- 通过实验展示了用合成图像训练得到的超分辨网络模型可以有效地处理真实图像复杂的退化类型。

论文链接：http://www4.comp.polyu.edu.hk/~cslzhang/paper/CVPR18_SRMD.pdf

CVPR2018 阿里巴巴 Poster 论文：基于尺度空间变换的本征图像分解

Intrinsic Image Transformation via Scale Space Decomposition

Lechao Cheng[1] Chengyi Zhang[1] Zicheng Liao[1,2]
College of Computer Science, Zhejiang University[1]
Alibaba - Zhejiang University Joint Institute of Frontier Technologies[2]

摘要

我们引入了一种新的网络结构，用于将图像分解为其本征的反射图像和光照图像。我们把它看作是一个图像到图像的转换问题，并且将输入和输出在尺度空间进行分解。通过将输出图像（反射图像和光照图像）扩展到它们的拉普拉斯金字塔的各个成分，我们开发了一种多通道网络结构，可以在每个通道内并行地学习到一个图像到图像转换函数，这个函数通过一个具有跳过连接的卷积神经网络来表示。该网络结构是通用的和可扩展的，并且已经在本征图像分解问题上表现出优异的性能。我们在两个基准数据集上评估了网络：MPI-Sintel 数据集和 MIT Intrinsic Images 数据集。定量和定性结果都表明，我们的模型在比之前最先进的技术上有了明显的进步。

1. 引言

最近在表示学习中出现了一种新兴的趋势，即学习从图像中分解出各个成分来解释输入的各个维度，例如照明，姿态和属性。然而，这个问题的初步形式之一，也就是分解图像为其本征的

反射图像和光照图像，并没有引起足够的重视。本征图像分解问题的解决方案将能够进行材料编辑，为深度估计提供线索，并为人类感知中长期存在的亮度恒定问题提供计算解释。然而，即使目前已经有了激动人心的进展，这个问题仍然是我们需要继续努力的一项艰巨任务。

部分难题来自这个问题的不确定性。基于反射图像和光照图像的先验知识，Retinex 算法将分解限制为梯度域中的阈值问题。这个模型有实用的效果，但不能处理复杂的材质或几何形状的锋利边缘，或是在强点光源下投射的阴影。另一部分难题在于图像渲染过程的复杂性，这是一个通过复杂的光学过程，将场景材质，几何和照明转换为二维图像的过程。本征图像分解的目的是试图部分地实现这个过程的逆向过程。

我们的工作中，我们使用深度神经网络作为函数逼近器来学习映射关系，从而在图像到图像的转换框架中处理本征图像分解过程。尽管已经提出了类似想法的模型（例如[37,32]），但我们的模型探讨了网络输入和输出的尺度空间，并且通过将函数逼近管线水平扩展为并行集的子带转换。

我们的工作的贡献主要在于，提出了一个用于本征图像分解的，基于尺度空间分解的生成网络。我们通过使用可学习的上/下采样器来建立经典的高斯和拉普拉斯金字塔结构来实现这个目的。我们最终的模型是一个复合网络，可以生成输出反射图像和光照图像的各个尺度下的分解;每个尺度下的分解由一个子网络预测，这些子网络的结果组合在一起成为我们的最终结果。

我们还提出了一个新的损失函数，它可以有效地保留图像的细节，保证反射图像的平滑和和光照的独特性质。我们进一步提出了一种数据增强的方式，以对抗标记数据的稀缺性 - 我们受到 Breeder 学习的启发，使用了一个预训练的网络给没有标签的图片生成标签，然后再对图片施加一定的扰动来生成新的数据集，最后用生成的数据集来提升我们网络的性能。

我们已经在 MPI-Sintel 数据集和 MIT 本征图像数据集上评估了我们的模型。实验结果证明了所提出的模型的有效性。我们的最终模型在各种评估指标上，与先前方法相比具有着显著的优势。

2. 相关工作（略）

3. 我们的方法

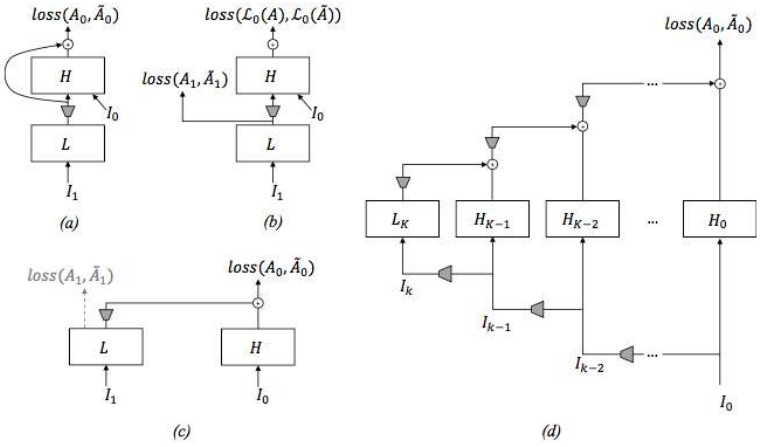
让我们首先考虑将输入图像 I 转换为输出图像 A，作为一个复杂的，高度非线性的和像素级的非局部映射函数 I→f（I）。已经很好地证明，深卷积神经网络是用于各种映射关系（从图像分类到图像到语言翻译）的通用和实用的参数化和优化框架。现在，让我们考虑如何使网络体系结构适应图像到图像转换问题，其中输入和输出都是具有自然细节层次（LOD）金字塔结构的图像，并且映射函数将输入链接到输出可能也会根据金字塔层次结构进行多通道分解。在下一节（3.1）

中，我们将描述从 ResNet 体系结构开始的模型改革过程，该体系结构将该属性用于我们的最终多通道分层网络体系结构。

我们将图像 I 的高斯金字塔写为[I0，I1，...，IK]，其中 I0 = I，K 是层的总数。我们用 Lk(I)= Ik - u(k + 1)表示第 k 个拉普拉斯金字塔层，其中 u 是上采样算子。根据定义，图像的拉普拉斯金字塔展开为 I = [L0(I)，L1(I)，...，Lk-1(I)，IK]，其中 L0(I)尺度和 IK 是高斯金字塔中定义的最低尺度层。

3.1 网络结构的演化

首先，让我们使用两个块（L 和 H）的简化网络来为低频带的映射建模映射 I→f（I）：L，并且 H 处理高频带中的映射以及任何残差这些网络由 L 省略。通过将 L 的输出与 H 的输出进行跳跃连接和求和，该网络是 ResNet 架构的实例。



接下来，通过在输出上应用拉普拉斯金字塔展开式，我们可以将（a）的损失分成两个分量：L 的输出被限制以适合低频高斯分量，而 H 的输出分别适合拉普拉斯细节分量（图 2-b）。这个改革后的网络等同于（a）但具有更严格的限制。

一个关键的过渡是从（b）到（c） - 因为通过将 L 的输出和 H 的输出连接起来，可以将两个堆叠的块重新连接成两个并联的分支，并且调整损失在 H 上相应地。所得到的网络结构（c）等价于（b） - 它们表示拉普拉斯分解方程的两种等价形式，即通过将剩余分量从左手边移到右手边并改变符号。 （c）中 L 的损失与正规化形式保持相同，我们的实验发现它是可选的。网络结构（d）是我们最终扩展模型的基础。

d，为此类似于拉普拉斯金字塔分解结构，我们为高频带引入多个子网络块 H0，H1，..... HK-1 和低频率的一个子网络块 LK：网络输入块级联下采样，并且网络块的输出被上采样并从左到右聚合以形成目标输出。在网络中学习下采样和上采样算子的所有参数（图 2 中的灰色阴影梯

形）。所有的网络块共享相同的架构拓扑结构，我们称之为“残差块”，并在第 3.2 节详细描述。

3.2 残差块

残余块是端到端卷积子网络，它们共享相同的拓扑结构，并将不同比例的输入变换为相应的拉普拉斯金字塔组件。每个残差块由 6 个连续级联的 Conv（3x3）-ELU-Conv（3x3）-ELU 子结构组成。由于我们预测输入图像的像素值，因此不使用完全连接的图层。我们采用在近期研究中流行的跳跃连接方案，其中包括 Huang 等人的 DenseNet 架构的一些变体。具体来说，在每个子结构中，最后一个 Conv 的输出按跳过连接进行每个元素的累加，结果是输入到最后一个 ELU 单元。中间层有 32 个特征通道，输出是 3 通道图像或残留图像。将 1x1 Conv 添加到第一层和最后一层的跳过连接路径中以进行尺寸扩展/缩减，以匹配剩余路径的输出。

我们使用指数线性单位（ELU）作为我们的激活函数，而不是使用 ReLU 和批量归一化，因为当 x ＜ 0 时 ELU 可以生成负激活值，并且具有零均值激活，这两者都可以提高对噪声的鲁棒性，当我们的网络变得更加深入时，训练的融合。此外，我们删除了 BN 层，因为它可以被 ELU 部分替换，ELU 速度提高了 2 倍，并且内存使用效率更高。

3.3 损失函数

我们的损失函数如下：

$$\mathcal{L} = \lambda_d \mathcal{L}_{data} + \lambda_p \mathcal{L}_{percep} + \lambda_t \mathcal{L}_{tv}$$

数据损失 数据损失定义了预测图像与真实之间的像素级相似性。 我们采用下面的联合双边滤波（也称为交叉双边滤波[13,39]），并结合预测的反照率和阴影的乘积应该与输入相匹配的约束，而不是使用像素方式的均方误差。

$$\mathcal{L}_{data} = \sum_{\mathcal{C}\in\{A,S\}} \frac{1}{N_p} \sum_{p\in\mathcal{C}} ||\mathcal{J}_p - \mathcal{C}_p||_2^2 + ||\tilde{A} * \tilde{S} - I||_2^2$$

$$\mathcal{J}_p = \frac{1}{\mathcal{W}_p} \sum_{q\in\mathcal{N}(p)} G_{\sigma_s}(||p-q||) G_{\sigma_r}(|\mathcal{C}_p - \mathcal{C}_q|) \tilde{\mathcal{C}}_p$$

$$\mathcal{W}_p = \sum_{q\in\mathcal{N}(p)} G_{\sigma_s}(||p-q||) G_{\sigma_r}(|\mathcal{C}_p - \mathcal{C}_q|)$$

感知损失 在转换过程中也应该保留高级语义结构，因此使用基于 CNN 特征的感知损失。 我们利用标准的 VGG-19 [44]网络从神经激活中提取语义信息。 我们的感知损失定义如下：

$$\mathcal{L}_{feat} = \sum_{\mathcal{C}\in\{A,S\}} \sum_{l} \frac{1}{F_l H_l W_l} ||\Phi_l(\tilde{\mathcal{C}}) - \Phi_l(\mathcal{C})||_2^2$$

变分损失 最后，我们使用变分项对输出结果进行平滑处理。

$$\mathcal{L}_{tv} = \sum_{\mathcal{C}\in\{A,S\}} \sum_{i,j} |\tilde{\mathcal{C}}_{i+i,j} - \tilde{\mathcal{C}}_{i,j}| + |\tilde{\mathcal{C}}_{i,j+1} - \tilde{\mathcal{C}}_{i,j}|$$

其中 i 和 j 是图像行和列索引。

3.4 数据增强训练

在本节中，我们将描述一种数据增强策略，用于将未标记的图像合并到自我增强网络培训过程中。我们从育种学习的工作中汲取灵感。这个想法是采用一个前向生成模型来为模型生成新的训练对，通过扰动模型产生的参数来增强。这种机制在一定程度上承载了 Boostrap 的精神，并且证明是相当有效的。例如，Li 等人最近将这种策略应用于外观建模网络中，通过根据模型对未标记图像的预测反射率生成训练图像。

我们从一个初步的网络开始，用一个中等大小的数据集进行训练，该数据集具有地面真实反射图像和光照图像。然后，我们将网络应用于一组新图像，并获得估计的反照率 A 和阴影 S。通过简单的综合程序，我们可以从估计中生成新图像。请注意，根据我们的损失定义，A 和 S 不会严格限制为与输入图像完全匹配，因此新合成的图像会偏离原始图像。

为了在增强数据集中引入进一步的扰动，我们另外应用自适应流形滤波器到 A 和 S，并使用过滤结果合成新数据。 AMF 滤波算子抑制 A 和 S 中可能来自输入图像或由早熟网络产生的噪声或不需要的细节，并用于调整新合成图像的多样性及其地面真实性。

4. 实验

在本节中，我们将描述 MPI-Sintel 数据集和 MIT Intrinsic Images 数据集上的模型评估。

4.1 数据集

MPI-Sintel 数据集由 18 个场景级计算机生成的图像序列组成，其中 17 个包含 50 个场景图像，一个包含 40 个图像。我们的实验中使用其中的 ResynthSintel 版本，因为数据满足 A×S = I 的约束条件。两种类型的训练/测试分割（场景分割和图像分割）用于与之前的工作进行逐一比较。场景分割将场景级别的数据集分割，其中一半场景用于训练，另一场景用于测试。图像分割会随机挑选一半图像进行训练/测试，而不考虑其场景类别。 原始版本的 MIT Intrinsic 数据集具有 20 个对象在实验室环境下的不同图像，每个图像具有 11 种不同的照明条件。

4.2 MPI 数据集实验结果

MPI-Sintel 数据集的评估结果如下图所示。同样，我们的模型比以前的方法产生了令人满意的结果，特别是在网络不易“过度拟合”测试数据的场景分割测试中。

与以前的工作比较：我们首先将我们的模型与以前的一系列方法进行比较，其中包括两个简单的基线 Constant Shading 和 Constant Albedo，一些传统方法以及最新的最新的基于神经网络的模型结果显示我们的模型无论是有/无数据增强训练，在所有三个指标的评估下，都有着最好的性能。

Sintel *image split*	si-MSE			si-LMSE			DSSIM		
	A	S	avg	A	S	avg	A	S	avg
Baseline: Constant Shading	5.31	4.88	5.10	3.26	2.84	3.05	21.40	20.60	21.00
Baseline: Constant Albedo	3.69	3.78	3.74	2.40	3.03	2.72	22.80	18.70	20.75
Color Retinex [18]	6.06	7.27	6.67	3.66	4.19	3.93	22.70	24.00	23.35
Lee et al. [30]	4.63	5.07	4.85	2.24	1.92	2.08	19.90	17.70	18.80
Barron & Malik [5]	4.20	4.36	4.28	2.98	2.64	2.81	21.00	20.60	20.80
Chen and Koltun [8]	3.07	2.77	2.92	1.85	1.90	1.88	19.60	16.50	18.05
Direct Intrinsic [38]	1.00	0.92	0.96	0.83	0.85	0.84	20.14	15.05	17.60
DARN [31]	1.24	1.28	1.26	0.69	0.70	0.70	12.63	12.13	12.38
Kim et al. [25]	0.7	0.9	0.7	0.6	0.7	0.7	9.2	10.1	9.7
Fan et al. [14]	0.67	0.60	0.63	0.41	0.42	0.41	10.50	7.83	9.16
Ours Sequential	0.83	0.74	0.79	0.58	0.54	0.56	7.61	7.91	7.76
Ours Hierarchical	0.81	0.78	0.79	0.58	0.58	0.58	8.18	7.16	7.62
Ours w/o Pyramid	0.92	1.37	1.15	0.65	1.15	0.90	8.44	10.96	9.70
Ours w/ MSE loss	0.72	0.62	0.67	0.62	0.46	0.50	7.98	6.37	7.18
Ours w/ ‘FPN’ input	0.73	0.60	0.67	0.49	0.43	0.46	6.84	6.76	6.80
Ours Final*	0.66	0.60	0.63	0.44	0.42	0.43	6.56	6.37	6.47
Ours Final+DA	**0.61**	**0.57**	**0.59**	**0.41**	**0.39**	**0.40**	**5.86**	**5.97**	**5.92**

Table 1. Quantitative Evaluation (×100) on the MPI-Sintel *image split*

Sintel *scene split*	si-MSE			si-LMSE			DSSIM		
	A	S	avg	A	S	avg	A	S	avg
Direct Intrinsic [38]	2.01	2.24	2.13	1.31	1.48	1.39	20.73	15.94	18.33
DARN [31]	1.77	1.84	1.81	0.98	0.95	0.97	14.21	14.05	14.13
Fan et al. [14]	1.81	1.75	1.78	1.22	1.18	1.20	16.74	13.82	15.28
Ours Sequential	1.61	1.56	1.58	1.05	1.11	1.08	10.24	11.90	11.07
Ours Hierarchical	1.59	1.51	1.55	0.98	1.01	0.99	8.70	9.55	9.13
Ours w/o Pyramid	1.82	2.01	1.92	1.01	1.39	1.20	14.43	14.27	14.35
Ours w/ MSE loss	1.47	1.44	1.46	0.92	0.95	0.93	9.48	10.97	10.23
Ours w/ ‘FPN’ input	1.46	1.40	1.43	0.96	0.97	0.97	8.50	9.30	8.90
Our Final*	1.38	1.38	1.38	0.92	0.93	0.92	8.46	9.26	8.86
Our Final+DA	**1.33**	**1.36**	**1.35**	**0.82**	**0.89**	**0.85**	**7.70**	**8.66**	**8.18**

Table 2. Quantitative Evaluation (×100) on the MPI-Sintel *scene split*

我们希望指出一个事实，所有方法对 Sintel 图像拆分的定量结果在某种程度上可能会产生误导。这是因为 Sintel 数据集中相同场景类别的图像序列彼此非常相似，所以通过在图像级别分割

所有数据（相同场景类型的图像可能出现在火车和测试集中），将全部训练集上的训练网络仍然会在图像分割测试集上“表现良好”。但场景分割数据集不会有这个问题。我们的实验中一个有趣的结果是，我们对前一结果的结果边缘在场景分割中比图像分割大。在表格中，尽管我们在图像分割上保留了相当适中的边缘，但我们在场景分割上保留的边距在 si-MSE 中高达 25%，在 DSSIM 中高达 43%，这表明我们网络结构在更具有挑战性的数据集上依然表现良好。

从顺序到并行结构 我们在第 3.1 节中描述的一个重要的网络架构改革是从顺序结构到多分支并行结构。这种改革将一个深度堆叠网络平滑作为一组并行通道，因此缓解了梯度反向传播问题，传播。该行（Ours Sequential）通过图 2 中的顺序架构（a）显示了结果。它显示该架构产生了与以前的作品相媲美的性能，但对最终的模型而言并不理想，特别是在 DSSIM 度量中。

分层优化 vs 联合优化 我们工作中的另一个架构优化是消除每个拉普拉斯金字塔等级的约束（损失），并同时训练所有网络通道并使用单个损失约束。在后一种情况下，我们称优化方案为联合优化，而前者为分级优化。补充材料中包括一些数值结果，解释层次优化的更多细节。在表 1-2 中，实验显示出在所有度量标准下联合优化方案有着 10%-15%的改进。

对其他因素的自我比较 我们也有一套关于其他因素的控制自我比较，包括金字塔结构，损失函数，交替网络输入和数据增加。

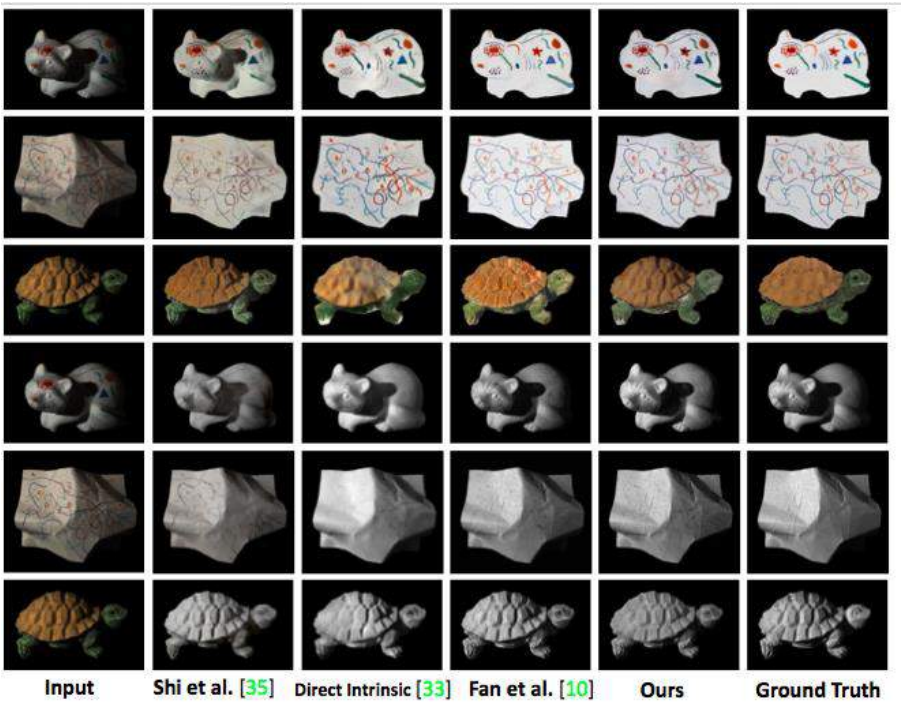
金字塔结构 实验（Ours w / o Pyramid）使用单通道网络显示结果，即，我们使用单个残差块直接从输入生成输出，而无须使用多通道分解结构。表 1 和表 2 中的结果表明，与对照设置相比，我们的金字塔结构的对应模型提高了 30%以上。请注意，随着金字塔层数的增加，网络复杂度呈线性增长直至一个常数因子。

损失函数 实验（Ours w / MSE loss）显示结果，用经典的 MSE 损失代替我们的损失函数。事实证明，MSE 损失的量化误差不会由于尺度不变 MSE 度量中的大因素而降低。然而，补充材料的定性结果确实揭示了 MSE 损失会产生模糊边缘的结果。基于结构的度量（DSSIM）在 MSE 损失和我们的损失之间也显示出更清晰的边际（从场景分割中的 10.23 到 8.86）。

CNN 特征作为输入 我们进一步研究在这个任务中高斯金字塔图像分量作为我们网络输入的影响，因为大多数现有的多尺度深度网络使用 CNN 网络产生的多尺度特征。实验（Ours w /’FPN’ input）显示将 CNN 特征完全作为 FPN 网络后面的输入的结果。比较显示我们的最终模型具有轻微但不明确的优势，这意味着 CNN 的高级特征仍然很好地保留了我们的像素到像素转换网络的大部分必要语义信息。

4.3 MIT 数据集实验结果

我们还在 MIT Intrinsic Images 数据集上评估了我们模型的性能，并与以前的方法进行了比较。结果如下图所示。在这组实验中，我们在两个不同的设置中进行了数据增强：Ours + DA 和 Ours + DA +。不同之处在于我们为增强所做的数据。 Ours + DA 是一种普通的设置，其中通过一组类似的对象类别名称从数据集中提取扩展数据。在 Ours + DA +中，我们在新的照明条件下。这会创建一个非常类似于原始数据集的数据集，在实际情况下几乎不可能获取。换句话说，它为增强数据的质量设定了一个上限。我们的结果表明，这两种增强设置都是有效的，而后者提供了我们可以从我们为此任务引入的数据增强方案中获得的限制线索。



Mit Intrinsic Data	si-MSE			LMSE
	Albedo	Shading	Average	Total
Zhou *et al.* [50]	0.0252	0.0229	0.0240	0.0319
Barron *et al.* [5]	0.0064	0.0098	0.0081	0.0125
Shi *et al.* [42]	0.0216	0.0135	0.0175	0.0271
Direct Intrinsic *et al.* [38]	0.0207	0.0124	0.0165	0.0239
Fan *et al.* [14]	0.0127	0.0085	0.0106	0.0200
Ours*	0.0089	0.0073	0.0081	0.0141
Ours + DA	0.0085	0.0064	0.0075	0.0133
Ours + DA$^+$	0.0074	0.0061	0.0068	0.0121

5. 结论

我们引入了受拉普拉斯金字塔启发的神经网络架构来进行本征图像分解。我们的神经网络将这个问题建模为不同尺度下的图像到图像的转换。我们在 MPI Sintel 和 MIT 数据集上进行了实验，实验结果表明我们的算法可以得到不错的数值结果和图片质量。对于未来的工作，我们期望所提出的网络架构能够在其他图像到图像转换问题上进行测试和改进，例如像素标记或深度回归。

CVPR2018 阿里巴巴 Poster 论文：基于直推式无偏嵌入的零样本学习

Transductive Unbiased Embedding for Zero-Shot Learning

Jie Song[1], Chengchao Shen[1], Yezhou Yang[2], Yang Liu[3], and Mingli Song[1]

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2]Arizona State University, Tempe, USA
[3]Alibaba Group, Hangzhou, China

摘要

大多数现有的零样本学习（Zero-Shot Learning，ZSL）方法都存在强偏问题：训练阶段看不见（目标）类的实例在测试时往往被归类为所看到的（源）类之一。因此，在广义 ZSL 设置中部署后，它们的性能很差。在本文，我们提出了一个简单而有效的方法，称为准完全监督学习（QFSL），来缓解此问题。我们的方法遵循直推式学习的方式，假定标记的源图像和未标记的目标图像都可用于训练。在语义嵌入空间中，被标记的源图像被映射到由源类别指定的若干个嵌入点，并且未标记的目标图像被强制映射到由目标类别指定的其他点。在 AwA2，CUB 和 SUN 数据集上进行的实验表明，我们的方法在遵循广义 ZSL 设置的情况下比现有技术的方法优越 9.3％至 24.5％，在遵循传统 ZSL 设置下有 0.2％至 16.2％的提升。

1. 引言

在大规模的训练数据集的支撑下，计算机视觉中的物体识别算法在近几年取得了突破性的进展。但是人工收集和标注数据是一项十分耗费人力物力的工作。例如，在细粒度分类中，需要专家来区分不同的类别。对于如濒临灭绝的物种，要收集到丰富多样的数据就更加困难了。在给定有限或者没有训练图片的情况下，现在的视觉识别模型很难预测出正确的结果。

零样本学习是一类可以用于解决以上问题的可行方法。零样本学习区分 2 种不同来源的类，**源类**(source)和**目标类**(target)，其中源类是有标注的图像数据，目标类是没有标注的图像数据。为了能够识别新的目标类（无标注），零样本学习假定源类和目标类共享同一个语义空间。图像和类名都可以嵌入到这个空间中。语义空间可以是属性(attribute)、词向量（word vector）等。在该假设下，识别来自目标类的图像可以通过在上述语义空间中进行最近邻搜索达成。

根据目标类的无标注数据是否可以在训练时使用，现有的 ZSL 可以分为 2 类：归纳式 ZSL(inductive ZSL)和直推式 ZSL(transductive ZSL)。对于归纳式 ZSL，训练阶段只能获取得到源类数据。对于直推式 ZSL，训练阶段可以获取到有标注的源类数据和未标注的目标类数据。直推式 ZSL 希望通过同时利用有标注的源类和无标注的目标类来完成 ZSL 任务。

在测试阶段，大多数现有的归纳式 ZSL 和直推式 ZSL 都假定测试图像都来源于目标类。因此，对测试图片分类的搜索空间被限制在目标类中。我们把这种实验设定叫作**传统设定**（conventional settings）。然而，在一个更加实际的应用场景中，测试图像不仅来源于目标类，还可能来自源类。这种情况下，来自源类和目标类的数据都应该被考虑到。我们把这种设定叫作**广义设定**（generalized settings）。

现有的 ZSL 方法在广义设定下的效果远差于传统设定。这种不良的表现的主要原因可以归纳如下：ZSL 通过建立视觉嵌入和语义嵌入之间的联系来实现新的类别的识别。在衔接视觉嵌入和语义嵌入的过程中，大多数现有的 ZSL 方法存在着**强偏**（strong bias)的问题（如图 1 所示）：在训练阶段，视觉图片通常被投影到由源类确定的语义嵌入空间中的几个固定的点。这样就导致了在测试阶段中，在目标数据集中的新类图像倾向于被分到源类当中。

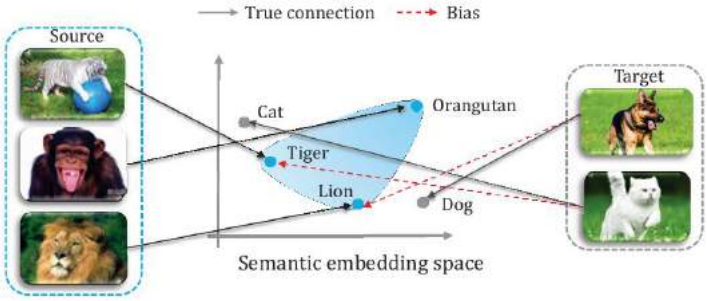


Figure 1. An illustrative diagram of the bias towards seen source classes in the semantic embedding space. The blue circles denote the anchor points specified by the source classes.

为了解决以上问题，本文提出了一种新的直推式 ZSL 方法。我们假定有标注的源数据和目标数据都可以在训练阶段得到。一方面，有标注的源数据可以用于学习图像与语义嵌入之间的关系。另外一方面，没有标注的目标数据可以用于减少由于源类引起的偏置问题。更确切地来说，我们的方法允许输入图像映射到其他的嵌入点上，而不是像其他 ZSL 方法将输入图像映射到固定的由源类确定的几个点上。这样有效地缓解了偏置问题。

我们将这种方法称为**准全监督学习（Quasi-Fully Supervised Learning, QFSL）**。这种方法和传统的全监督分类工作方式相似，由多层神经网络和一个分类器组成，如图 2 所示。神经网络模型架构采用现有的主流架构，比如 AlexNet、GoogleNet 或者其他框架。在训练阶段，我们的模型使用有标注的源类数据和没有标注的目标数据进行端到端的训练。这使得我们的模型有一两个个明显的特性：（1）如果未来可以得到目标类的标注数据，那么标注数据可以直接用于进一步训练和改进现有的网络模型；（2）在测试阶段，我们得到的训练模型可以直接用于识别来自于源类和目标类的图像，而不需要进行任何修改。

本论文的主要贡献总结如下：

- 提出了准全监督学习的方法来解决零样本学习中的强偏问题。据我们所知，这是第一个采用直推式学习方法来解决广义设定下零样本学习问题。
- 实验结果表明我们的方法在广义设定下和传统设定下都远超现有的零样本学习方法。

2. 相关工作（略）

3. 我们的方法

3.1 问题的形式化

假设存在一个源数据集$\mathcal{D}^s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$，每张图片$x_i^s$与相应的标签$y_i^s$对应，其中$y_i^s \in \mathcal{Y}^s = \{(y_i)\}_{i=1}^{S}$，S表示源类中类的个数。目标数据集$\mathcal{D}^t = \{(x_i^t, y_i^t)\}_{i=1}^{N_t}$，每张图片$x_i^t$与相应的标签$y_i^t$对应，其中$y_i^t \in \mathcal{Y}^t = \{(y_i)\}_{i=1}^{T}$，T表示目标类中类的个数。ZSL 的目标就是学习如下所示的预测函数$f(\cdot)$:

$$f(x; W) = \text{argmax}_{y \in \mathcal{Y}} F(x, y; W),$$

其中$F(\cdot)$是一个得分函数，其目标是正确的标注比其他不正确的标注具有更高的得分。W是模型$F(\cdot)$的参数，$F(\cdot)$通常使用如下的双线性形式：

$$F(x, y; W) = \theta(x)^T W \Phi(y),$$

其中$\theta(x)$和$\Phi(y)$分别表示视觉嵌入和语义嵌入。得分函数通常使用带正则化的目标函数进行优化：

$$\mathcal{L} = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathcal{L}_p(y_i, f(x_i; W)) + \gamma\Omega(W)$$

其中$\mathcal{L}_p$表示分类损失，用于学习视觉嵌入和语义嵌入之间的映射。Ω表示用于约束模型复杂度的正则项。

本文假设给定标注源数据集$\mathcal{D}^s$，无标注目标数据集$\mathcal{D}_u^t = \{(x_i^t)\}_{i=1}^{N_t}$和语义嵌入$\Phi$，学习 ZSL 模型，使得其既能在传统设定下又能在广义设定下获取良好的表现。

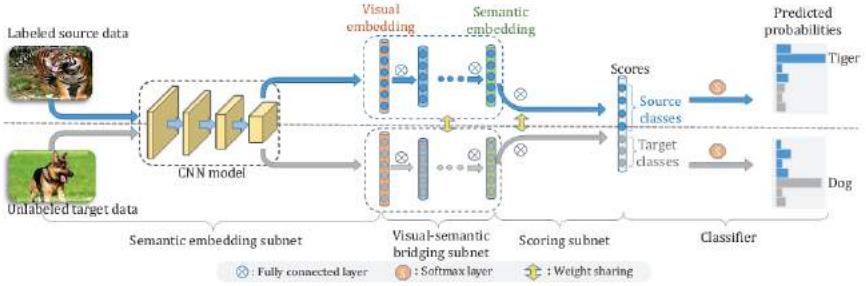


Figure 2. An overall architecture of the proposed QFSL model. Both the labeled and the unlabeled data are used to train the same model. Here for a better understanding, we depict them in two streams.

3.2 QFSL 模型

不同于以上描述的双线性形式，我们将得分函数F设计成非线性形式。整个模型由深度神经网络实现。模型包括 4 个模块：视觉嵌入子网络，视觉-语义衔接子网络，得分子网络和分类器。视觉嵌入子网络将原始图像映射到视觉嵌入空间。视觉-语义衔接子网络将视觉嵌入映射到语义嵌入子网络。得分子网络在语义空间中产生每一类的得分。分类器根据得分输出最终的预测结果。所有的模块都是可微分的，包括卷积层，全连接层，ReLU 层和 softmax 层。因此，我们的模型可以进行端到端的训练。

视觉嵌入子网络

现有的大多数模型采用了 CNN 提取得到的特征作为视觉嵌入。在这些方法中，视觉嵌入函数θ是固定的。这些方法并没有充分利用深度 CNN 的强大的学习能力。本文采用了预训练的 CNN 模型来进行视觉嵌入。我们的视觉嵌入模型的主要不同之处在于可以和其他模块一起进行优化。视觉嵌入模块的参数记为W_θ。除非特别说明，我们把第一个全连接层的输出作为视觉嵌入。

视觉-语义衔接子网络

衔接图像和语义嵌入之间的关系对 ZSL 来说很重要。这种关系可以通过线性函数或者非线性函数来建模。本文采用了非线性函数φ将视觉嵌入映射到语义嵌入。φ由若干个全连接层来实现，其中每一个全连接层后面跟了一个非线性激活函数：ReLU。衔接函数的设计依赖于上述的视觉嵌入子网络的架构。具体来说，我们的设计是按照所选择 CNN 模型的全连接层来设计的。视觉-语义衔接子网络和视觉嵌入网络一起进行优化。视觉-语义衔接子网络参数记作W_φ。

得分子网络

衔接视觉嵌入和语义嵌入之后，识别任务可以通过在语义嵌入空间中使用最近邻搜索来实现。

给定一张图像，我们首先通过视觉嵌入子网络得到它的视觉嵌入。然后，利用视觉-语义衔接子网络，完成从视觉嵌入到语义嵌入的映射。最后，我们通过内积计算得到投影得到的视觉嵌入和语义嵌入的得分。因此，得分函数可以表示如下：

$$F(x,y;W)=\varphi(\theta(x;W_\theta);W_\varphi)\Phi^*(y)$$

其中W_θ和W_φ分别为视觉嵌入函数和视觉-语义衔接函数的权重，$\Phi^*(y)$是y的归一化语义嵌入：$\Phi^*(y)=\frac{\Phi(y)}{\|\Phi\|_2}$。

得分函数由单个全连接层来实现。它的权重使用源类和目标类的归一化语义：$[\Phi^*(y_1),\Phi^*(y_2),\cdots,\Phi^*(y_{S+T})]$来初始化。和视觉嵌入子网络和视觉-语义衔接子网络不同的是，得分子网络的权重是固定的，在训练阶段不参与更新。通过这种方式，我们的模型将图像x_i^s投影到与视觉嵌入$\Phi^*(y_i^s)$相近的方向上。

需要注意的是目标类的数据没有标注，这些数据在我们的方法中用到了训练阶段当中。因此，在训练阶段，我们的模型对于一张给定的图像，产生了$S+T$个得分。

分类器

经过得分函数后，我们使用$(S+T)$路的 softmax 分类器产生了所有类的概率。输入图像的预测结果为概率最高的那个类。

3.3 模型优化

我们的方法采用了类似于由$(S+T)$路的 softmax 分类器的全监督分类模型，用来分类目标类和源类。但是，只有源类数据是有标注的，目标类数据没有标注。我们定义了准全监督损失函数来训练提出的模型：

$$\mathcal{L}=\frac{1}{N_s}\sum_{i=1}^{N_s}\mathcal{L}_p(x_i^s))+\frac{1}{N_t}\sum_{i=1}^{N_t}\lambda\mathcal{L}_b(x_i^t))+\gamma\Omega(W)$$

通常，传统的全监督分类器的损失函数包括分类损失$\mathcal{L}_p$和正则化损失Ω。和传统定义不同，我们提出的 QFSL 结合了一个额外的偏置损失$\mathcal{L}_b$来缓解强偏问题：

$$\mathcal{L}_b(x_i^t)=-ln\sum_{i\in y^t}p_i$$

其中，p_i表示预测为类i的概率。给定一个来自目标类的实例，该损失鼓励模型增加所有目标类的概率和。这样可以防止目标类被映射到源类中。

对于分类损失$\mathcal{L}_p$，我们采用了交叉熵。对于正则化损失Ω，L_2范数来约束训练参数$W=\{W_\theta,W_\varphi\}$。λ和γ用于平衡不同损失之间的权重，通过交叉验证来确定。在训练阶段，所有标注的数据和未标注的数据混合在一起作为训练数据。模型使用随机梯度下降算法(SGD)进行优化。每一个批(batch)训练图像从混合数据集中随机抽取。实验结果表明我们的方法不仅有效地避免了偏置问题，还帮助建立起了更好的视觉嵌入和语义嵌入之间的联系。

4. 实验

4.1 数据集

我们在三个数据集上评估了我们的方法。这三个数据集分别为 AwA2， CUB， SUN。在实验中，我们采用属性作为语义空间，用类平均准确度衡量模型效果。

4.2 在传统设置下的效果比较

首先我们在传统设置下对我们方法和现有方法。用来做对比的现有方法分为两类：一类是是归纳式方法，包括 DAP，CONSE，SSE，ALE，DEVISE，SJE，ESZSL，SYNC；另一类是直推式方法，包含 UDA，TMV，SMS。与此同时，还比较了一个潜在的 baseline（标记为 QFSL$^-$）：只用有标注的源数据来训练我们的模型。实验效果如表 1。可以看出，我们的方法大幅度（4.5~16.2%）提升了分类准确度。

	Method	CUB SS	CUB PS	SUN SS	SUN PS	AwA2 SS	AwA2 PS
§	DAP [19]	37.5	40.0	38.9	39.9	58.7	46.1
	CONSE [25]	36.7	34.3	44.2	38.8	67.9	44.5
	SSE [43]	43.7	43.9	25.4	54.5	67.5	61.0
	ALE [1]	53.2	54.9	59.1	58.1	80.3	62.5
	DEVISE [9]	53.2	52.0	57.5	56.5	68.6	59.7
	SJE [2]	55.3	53.9	57.1	53.7	69.5	61.9
	ESZSL [31]	55.1	53.9	57.3	54.5	75.6	58.6
	SYNC [4]	54.1	55.6	59.1	56.3	71.2	46.6
£	UDA [16]	39.5	–	–	–	–	–
	TMV [10]	51.2	–	61.4	–	–	–
	SMS [12]	59.2	–	60.5	–	–	–
	QFSL$^-$	58.5	58.8	58.9	56.2	72.6	63.5
		↑10.5	↑13.3	↑0.3	↑0.2	↑4.5	↑16.2
	QFSL	**69.7**	**72.1**	**61.7**	**58.3**	**84.8**	**79.7**

表 1. 在传统设置下的实验比较

4.3 在广义设置下的效果比较

大多数现有直推式方法在测试阶段都采用了同训练阶段同样的数据来评估性能。然而，如果我们的方法也采用这种方式来评估效果是很不合理的。因为我们的方法已经利用到了无标签的数据来源于目标类这一监督信息。为了解决这一问题，我们将目标数据平分为两份，一份用来训练，另一份用来测试。然后交换这两份数据的角色，再重新训练一个模型。最终的效果为这两个模型的平均。我们比较了我们的方法和若干现有方法，以及一个隐含的 baseline：先训练一个二分类器来区分源数据和目标数据，然后再在各自搜索空间中分类。实验结果如表 2。

	Method	AwA2 MCA_s	AwA2 MCA_t	AwA2 H	CUB MCA_s	CUB MCA_t	CUB H	SUN MCA_s	SUN MCA_t	SUN H
†	DAP [19]	84.7	0.0	0.0	67.9	1.7	3.3	25.1	4.2	7.2
	CONSE [25]	90.6	0.5	1.0	72.2	1.6	3.1	**39.9**	6.8	11.6
	SSE [43]	82.5	8.1	14.8	46.9	8.5	14.4	36.4	2.1	4.0
	ALE [1]	81.8	14.0	23.9	62.8	23.7	34.4	33.1	21.8	26.3
	DEVISE [9]	74.7	17.1	27.8	53.0	23.8	32.8	30.5	14.7	19.8
	SJE [2]	73.9	8.0	14.4	59.2	23.5	33.6	30.5	14.7	19.8
	ESZSL [31]	77.8	5.9	11.0	63.8	12.6	21.0	27.9	11.0	15.8
	SYNC [4]	90.5	10.0	18.0	70.9	11.5	19.8	43.3	7.9	13.4
	CMT [36]	90.0	0.5	1.0	49.8	7.2	12.6	21.8	8.1	11.8
‡	CMT* [36]	89.0	8.7	15.9	60.1	4.7	8.7	28.0	8.7	13.3
	CS [6]	*77.6*	*45.3*	*57.2*	*49.4*	*48.1*	*48.7*	*22.0*	*44.9*	*29.5*
	baseline	*72.8*	*52.1*	*60.7*	*48.1*	*33.3*	*39.4*	*18.5*	*30.9*	*23.1*
‡	*QFSL*^{G}	*92.4*$^{\uparrow 1.8}$	*64.3*$^{\uparrow 12.2}$	*75.8*$^{\uparrow 15.1}$	*74.2*$^{\uparrow 2.0}$	***71.6***$^{\uparrow 23.5}$	*72.9*$^{\uparrow 24.2}$	*33.6*$^{\downarrow 8.3}$	***54.8***$^{\uparrow 9.9}$	***41.7***$^{\uparrow 12.2}$
	QFSL^R	***93.1***$^{\uparrow 2.5}$	***66.2***$^{\uparrow 14.1}$	***77.4***$^{\uparrow 16.7}$	***74.9***$^{\uparrow 2.7}$	*71.5*$^{\uparrow 23.4}$	***73.2***$^{\uparrow 24.5}$	*31.2*$^{\downarrow 8.7}$	*51.3*$^{\uparrow 6.4}$	*38.8*$^{\uparrow 9.3}$

可以看出，我们模型的整体性能（调和平均数 H）有着 9.3~24.5 的明显提高。该项指标的提高主要得益于在目标数据上的效果提升，同时又没有在源数据上大幅度降低准确度。该结果表明，我们的方法能够很大程度上缓解强偏问题。

5. 讨论

现实世界中，目标类的数量可能远远高于源类数量。然而，大多数现有 ZSL 数据集的源、目标数据划分都违背了这一点。比如，在 AwA2 中，40 个类用来做训练，10 个类用来做测试。我们

在实验上给出了随着源数据类别的增加，QFSL 在效果上如何变化。该实验在 SUN 数据集上进行，72 类作为目标类，随机选取剩下的类作为源类。我们尝试了 7 个大小不同的源类集，类的数量分别为{100, 200, 300, 450, 550, 600, 645}。用这些不同大小的源类作为训练集，测试我们的方法，效果如图 3。由图可以看出，随着类别增加，模型能够学习到更多的知识，其在目标数据集上准确度越来越高。同时，由于源数据和目标数据变得越来越不平衡，强偏问题越来越严重。我们方法能够缓解强偏问题，因而其在效果上的优越性也越来越明显。

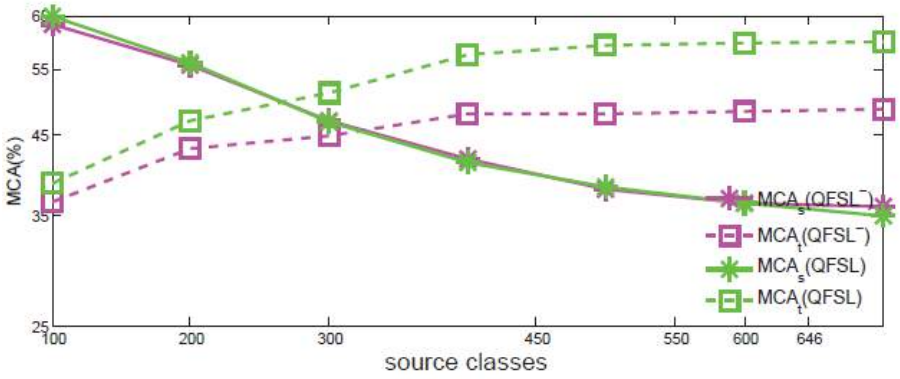


图 3. 准全监督在 SUN 数据集上效果

6. 结论

本文提出了一种用于学习 ZSL 无偏嵌入的直接但有效的方法。这种方法假设标注的源数据和未标注的目标数据在模型训练的过程中可以使用。一方面，将标注的源数据映射到语义空间中源类对应的点上。另外一方面，将没有标注的目标数据映射到语义空间中目标类对应的点上，从而有效地解决了模型预测结果向源类偏置的问题。在各种基准数据集上的实验表明我们的方法在传统设定和广义设定下，大幅超过了现有的 ZSL 方法。

论文原文地址：http://arxiv.org/abs/1803.11320

Alibaba Group
阿里巴巴集团

阿里技术

扫一扫二维码图案，关注我吧

「阿里技术」微信公众号

「阿里巴巴机器智能」微信公众号