论文推荐：Deep Transfer Learning for Multiple Class Novelty Detection (CVPR 2019 accepted paper)

潘皓
21821147
flewsoo@163.com

摘要（Abstract）

这篇论文提出了一种新的基于迁移学习的多类别新奇检测 (Multiple Class Novelty Detection)方法。该方法使用了一个端到端的深度学习结构，并利用了外部的、分布外的数据集来提升视觉新奇检测的深度网络的性能。论文中的方法和标准的深度分类网络主要有两个地方不同：第一，这篇论文在交叉熵损失函数之外引入了一个新的损失函数，成员损失(membership loss)。第二，论文中引入了一个新的机制，称为全局负滤波器(globally negative filter)，它能够使模型更好的利用外部数据。通过这两个改变，根据激活的阈值来进行新奇判断变得高效，实证表示这篇论文的方法在四个公开数据集上的表现均在时下先进的方法的基础上实现了明显的性能提升。

1. 引言

在计算机视觉中，图像识别是一个很重要的任务。近年来随着深度神经网络的发展，利用深度神经网络来进行图像识别引起了大量的关注。一个图像识别系统可以通过在训练中观察大量图片实例和它们的标签，抽取其中可用于判断的特征和模式。一个实用的图像识别系统需要首先判断观察的实例是否属于给定的已知类别的集合，如果属于则进而判断其属于哪一个类别。

这篇论文[1]提出了一种新的基于迁移学习的多类别新奇检测 (Novelty Detection)方法。该方法使用了一个端到端的深度学习结构，并利用了外部的、分布外的数据集来提升视觉新奇检测的深度网络的性能。

判断一个观察的实例是否属于给定的已知类别被称为新奇检测[2] (Novelty Detection)。在实际训练过程中，对于给定已知类别的图像数据集，通常不会有相应的未知类别的数据集，但是可以引入在其它问题领域的不属于给定已知类别的图像数据集，这些数据被称为分布外(out-of-distributional) 样本。比如在狗的种类识别应用中，ImageNet数据集就可以引入进来作为分布外样本。一个理想的分布外样本是随机抽取除了给定已知类别的图像外的所有自然图像，但是由于可获得的分布外样本通常是来自于其它问题领域的样本，它们和自然图像样本还是有所差别的。

尽管分布外样本不完全是自然的，但是由于深度神经网络会提取出一些可泛化的抽象特征，这些从分布外样本中提取的知识还是可以迁移到原来的问题中帮助进行新奇检测。

相对于传统的深度网络新奇检测方法，这篇论文介绍的方法的亮点主要有四点：

1. 这篇论文中提出的方法是端到端的的深度学习结构，这是首次在新奇检测问题上应用端到端的深度学习结构。

2. 引入了一个新的损失函数，成员损失 (membership loss)，这个损失函数功能类似于交叉熵损失函数，并且可以鼓励已知类别的实例产生更高的激活。

3. 提出利用大规模外部数据来学习一个新的机制，称为全局负滤波器 (globally negative filter)，当观察图像是新奇图像时，它可以抑制图像在已知类别上产生高的激活值。

4. 这篇论文提出的方法在性能表现上明显优于当前的先进方法。

2. 相关工作

新奇检测的方法有很多，一种最简单的方法是把所有给定的已知类别当作一个类别，使用one-class SVM[3]或者SVDD[4]进行分类，得到一个分割已知类别样本的超平面，这种方法的缺点是性能表现较差。

一种新的方法是KNFST[5](Kernel Null Space-based Novelty Detection)，这种方法把原来的样本映射到一个新的子空间（称为null 空间），在这个子空间种，不同类别的样本会分别映射到对应类别的一个单独的点上，即在这个子空间中类内的方差是 0，于是待检测的图像只需要度量该图像映射后的点与代表已知类别的点的最小距离，就可以判断这个图像是否是新奇图像。非核方法的NFST只能应用于小样本的分析中，而引入了核方法后则可以处理更大的样本。KNFST在非深度学习的方法中的性能表现是最好的。

深度学习的方法在新奇检测中比传统方法们表现得更好。目前先进的方法是引入分布外样本，利用从分布外样本中获取的知识迁移到原来的问题中帮助进行新

奇检测。通常的做法[6]是将网络在分布外样本上进行预训练，然后将预训练后的样本在原来的样本上进行微调，也可以将原来的样本和分布外样本一起进行微调。最后取定一个阈值，如果每个类别的激活分数都低于这个阈值，则认为被检测的样本是新奇的。

3. 模型介绍

这篇论文首先在传统的深度卷积网络的结构上提出了正、负滤波器的概念。正、负滤波器分别对图像是否属于特定类别提供正向、反向的支持。然后探讨了交叉熵损失函数的局限性，这种局限性使模型在正确类别上获得一个低的激活值时不能够得到足够的惩罚，这就使正确类型的激活值偏低，在新奇检测时造成更多的假负例(FN, False Negative)；同时当正确类别的确获得了最高的激活值时，其它非正确类别获得高激活值的现象也没有获得相应的惩罚。这就反映了模型缺少机制来抑制错误类别的高激活，在新奇检查时造成更多的假正例(FP, False Positive)。因此在文中引入了一个新的成员损失函数和一个新的全局负滤波器机制来解决这些弊端。

3.1. 正、负滤波器

这篇论文中提出的模型是基于迁移学习的深度学习架构上的端到端的新奇检测模型。论文首先提出了正滤波器 (positive filters)和负滤波器(negative filters)的概念。一个标准的深度神经网络会接受一个图像输入并且产生一个激活。考虑一个 c 类别的有监督图像分类问题，训练样本为 $\mathbf{x} = x_1, x_2, \ldots, x_n$，其相应的标签为 $\mathbf{y} = y_1, y_2, \ldots, y_n$，其中 $y_i \in \{1, 2, \ldots, c\}$。深度卷积网络试图学习一个层次的、卷积的网络结构，不同层次的卷积核会提取图像实例输入的不同级别的抽象，在 c 类别的分类中，最顶层的卷积层的输出 **g**（论文中称为激活）会经过一系列线性和非线性的变幻来产生最后的激活向量 $\mathbf{f} \in R^c$（比如，**g** 是VGG16 中的conv5-3 层，Resnet50 中的conv5c层。**f** 是相应网络中的fc8 层和fc1000 层）。在有监督条件下，网络试图学习参数使得对 $\forall i \in \{1, 2, \ldots, n\}$，都满足 $\text{argmax}\, \mathbf{f} = y_i$。这通常是通过基于交叉熵损失函数来训练网络的参数。

假如最顶层的卷积层有 k 个卷积核，其输出 **g** 是 k 个激活图。最后激活向量 **f** 是 **g** 的一个函数。对于给定的类别 i，会存在 k_i 个卷积核 （$1 \leqslant k_i \leqslant k$）会产生正的激活值。这些正的激活提供了观察图像是来自类别 i 的支持。反之，同样存在会产生负的激活值的卷积核，其提供依据来反对图像是来自类别 i 的。这些对特定类别提供正向或反向支持的卷积核被称为正滤波器(positive filters)和负滤波器(negative filters)。

以Resnet结构作为例子，其最顶层卷积层的输出 **g** 首先经过一个全局平均池化，紧跟着接着一个全连接层。于是最后激活向量 **f** 的第 i 个分量 $\mathbf{f}_i$ 可以被写成 $\mathbf{f}_i = W_i \times \text{GAP}(\mathbf{g})$，其中GAP是全局平均池化操作，W 是全连接层的权重矩阵。因此，第 i 个类别的激活是最顶层卷积层的输出 **g** 的加权和。那些在 W 中权重为正的卷积核称为正滤波器，相应的权重为负的卷积核称为负滤波器。

考虑ILRVRC12 中的沙蛇类别。见 Figure 1，上方是全连接层输入对于沙蛇类别的不同权重。把权重为正的卷积核称为正滤波器，相应的权重为负的卷积核称为负滤波器。下方是正、负滤波器的可视化图像。这些模式倾向于对产生高的激活。可以看到最高的几个正滤波器会被类似于蛇的结构激活，另一方面，最高的几个负滤波器则与沙蛇的外观毫无关联。

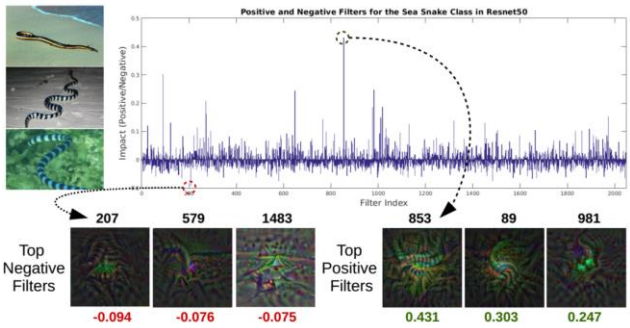

Figure 1. 沙蛇类别在Resnet50 中的正、负滤波器，训练于ILSVRC12 数据集上。上方：全连接层输入对于沙蛇类别的不同权重。下方：正、负滤波器的可视化图像。这些模式倾向于对沙蛇的图像产生高的激活。可以看到最高的几个正滤波器会被类似于蛇的结构激活。（图片来自原论文）

3.2. 成员损失

在分类神经网络中，激活向量 **f** 会最后经过一个softmax层产生一个归一化的激活向量 $\tilde{f}$，其每个分量可以被看做图像属于对应类别的可能性大小。神经网络通过最小化交叉熵损失函数，同时也是其负的对数似然函数来学习这一机制。但是由于经过softmax后的向量体现的是一个相对的度量，最终损失函数的损失只考虑了激活向量 f 中正确类别分量相对于其它类别分量的相对大小，而没有考虑其激活值的绝对大小。这就导致了即使该图像对于正确类别的激活值很低，只要其它类别的激活值更低，也能得到一个很小的损失函数。当进行分类问题时，神经网络依旧能预测出正确的类别，但是当进行新奇检测时，由于判断新奇与否是通过设定一个激活值阈值，低的激活值可能会错误的预测其是新奇图像。

另一方面，交叉熵损失函数缺少对无关类别的高激活的惩罚，这就使一些不准确的类别之间的关系容易在训

练中被错误的学习进来。

于是这篇文章的作者引入了成员损失作为损失函数的补充。在多分类问题中，在最后一层通常会使用一个softmax函数来进行归一化，归一化后向量的各个分量被视作图像属于相应类别的概率。这篇在和softmax层平行的地方加入了一个新的sigmoid层，其输入和softmax层入相同。sigmoid层的输出各分量的总和不是 1，其输出不是一个概率分布，但是可以看作是图像属于各类别概率大小的一种表示。如果y=i,我们显然希望sigmoid的输出的第i个分量趋近于 1，反之则趋近于 0。

于是引入成员损失，这个损失函数希望使已知类别的sigmoid输出对应的分量趋近于 1，其它则趋近于 0。

其总的公式为：

$$L_M(\mathrm{x},y)=[1-\sigma(f(x)_y)]^2+\lambda\frac{1}{c-1}\sum_{i=1,i\neq y}^{c}[\sigma(f(x)_i]^2$$

左边的式子表示若y = i,我们显然希望sigmoid的输出的第i个分量趋近于 1，右边的式子则希望其它分量的输出值趋近于 0。λ控制两倍的权重，推荐的取值为λ=5。利用成员损失函数，每中类型单独的激活值将会在绝对而不仅仅是相对情况下被考虑。低激活值的已知类型的样本会被惩罚，即使其它类型的激活值也很低。当成员损失函数和交叉熵损失函数配合使用时，模型会学到一些结构，使正确的类型得到更高的激活值。

3.3. 全局负滤波器

当使用传统的分类网络时，新奇的图片经常会产生高的激活值，导致假正例错误的发生。这是由于一些新奇的图片会激活一些正滤波器，但是却没有激活负滤波器来抵消正滤波器的影响。这是因为网络学习到的对一个类型的负滤波器，实际上只是另一个类型的高权重的正滤波器。当这些当前类别的负滤波器产生高激活值时，实际上是网络判断观察的图像是属于另一种已知类别，而不是网络判断了观察的图像不属于当前类别，网络实际没有学习出专门判断图像是否是新奇的结构。

所以这篇论文的作者认为需要学习一些卷积核（滤波器），它们对于所有已知类型来说都是负滤波器。这些滤波器被称为全局负滤波器，当它们产生高激活值时，将会有效的支持观察的图像是新奇图像的判断。

为了学习到全局负滤波器，文中引入一个联合学习(joint-learning)的网络结构。利用分布外样本加入训练中，假如分布外样本包含 C 个类别，那么网络一共会学得 $c + C$ 个类别的正滤波器，而对应于分布外样本的正滤波器可以近似的起到全局负滤波器的作用。分布外样本规模越大、分布越离散，学习到的全局负滤波器效果越好。

3.4. 训练和测试过程

首先引入一个联合学习的网络结构，见Figure 2 的左半部分。首先选择一种深度卷积神经网络作为骨架（可以是简单的网络比如Alexnet，或者复杂的如DenseNet），见图中的蓝色部分。两个并行的相同CNN骨架一起进行训练，这两个CNN结构共享参数，所以也可以看出同一个CNN骨架被复用。训练过程中，图上方的CNN结构使用分布外样本R作为输入（文中称作

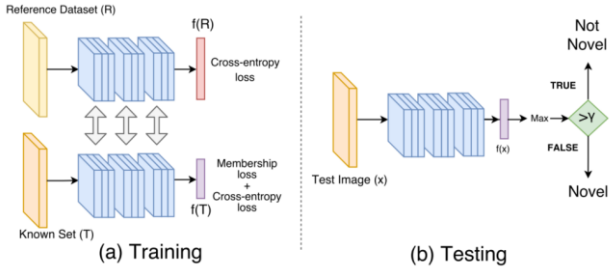

Figure 2. 左图（a）是训练的模型，右图（b）是左图的下半部分，用于最后的预测。（图片来自原论文）

Reference Dataset），图下方的CNN使用已知类型的数据集 T（需要识别的类型的数据）作为输入。两个CNN结构通过一个全连接层与输出向量连接，输出向量的维度是其输入数据的类别大小，两个部分的全连接层参数是独立的。在上方的使用样本外数据 R 训练的网络结构中，使用传统的交叉熵损失函数，而在下方使用已知样本 T 训练的网络中，使用成员损失函数和交叉熵损失函数的线性加权作为最终的损失函数，默认是取两者平均。

在测试过程中，丢弃训练网络钟的上半部分，通过选取一个阈值，判断最终输出的最大激活值是否超过这个阈值来进行新奇判断。

4. 实验结论

这篇论文的作者在VGG16和AlexNet两种网络的基础上分别进行了四个公开数据集Caltech-256、CUB-200、Dogs和FounderType数据集上进行了检验。对比了其与One-class SVM、传统的两种微调(Finetune)的迁移学习网络、多种KNFST以及OpenMax方法下的AUC数据。实证表明，文中提出的方法对比时下先进的方法取得了显著的提升。

消融实验的结果：

a) 单个CNN结合交叉熵损失函数，AUC为 0.854。

b) 单个CNN结合成员损失函数和交叉熵损失函数加权，AUC为 0.865。

Table 1. 在四个数据集上的新奇检测的结果(AUC of ROC curve)。

Method	Caltech-256		CUB-200		Dogs		FounderType	
	VGG16	AlexNet	VGG16	AlexNet	VGG16	AlexNet	VGG16	AlexNet
Finetune[20], [9]	0.827	0.785	0.931	0.909	0.766	0.702	0.841	0.650
One-class SVM[17]	0.576	0.561	0.554	0.532	0.542	0.520	0.627	0.612
KNFST pre[3]	0.727	0.672	0.842	0.710	0.649	0.619	0.590	0.655
KNFST[3], [10]	0.743	0.688	0.891	0.748	0.633	0.602	*0.870*	0.678
Local KNFST pre[2]	0.657	0.600	0.780	0.717	0.652	0.589	0.549	0.523
Local KNFST[2]	0.712	0.628	0.820	0.690	0.626	0.600	0.673	0.633
K-extremes[18]	0.546	0.521	0.520	0.514	0.610	0.592	0.557	0.512
OpenMax[1]	0.831	0.787	*0.935*	*0.915*	0.776	*0.711*	0.852	0.667
Finetune($c+\mathcal{C}$)	*0.848*	*0.788*	0.921	0.899	*0.780*	0.692	0.754	*0.723*
Deep Novelty (ours)	**0.869**	**0.807**	**0.958**	**0.947**	**0.825**	**0.748**	**0.893**	**0.741**

c) 两个并行CNN结合交叉熵损失函数，AUC为 0.864。
d) 完整的模型，AUC为 0.906。
可见成员损失函数和新的训练机制确实起了重要的作用。

5. 总结

这篇文章提出了一种新的基于迁移学习的多类别新奇检测方法。该方法使用了一个端到端的深度学习结构，并利用了外部的、分布外的数据集来提升视觉新奇检测的深度网络的性能。

1. 这篇论文中提出的方法是端到端的的深度学习结构，这是首次在新奇检测问题上应用端到端的深度学习结构。

2. 引入了一个新的损失函数成员损失 (membership loss)作为交叉熵损失函数的补充，这个损失函数功能类似于交叉熵损失函数，并且可以鼓励已知类别的实例产生更高的激活值。

3. 提出利用大规模外部数据来学习全局负滤波器(globally negative filter)，当观察图像是新奇图像时，它可以抑制图像在已知类别上产生高的激活值。

4. 这篇论文提出的方法在性能表现上明显优于当前先进的其它方法。

References

[1] Perera P , Patel V M . Deep Transfer Learning for Multiple Class Novelty Detection[J]. 2019.

[2] M. Markou and S. Singh. Novelty detection: a review – part1: statistical approaches. *Signal Processing*, 83(12):2481 –2497, 2003.

[3] B. Scholkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, ¨and R. C. Williamson. Estimating the support of a high dimensional distribution. *Neural Comput.*, 13(7):1443–
1471, 2001.

[4] D. M. J. Tax and R. P. W. Duin. Support vector data description. *Mach. Learn.*, 54(1):45–66, 2004.

[5] P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, andJ. Denzler. Kernel null space methods for novelty detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[6] A. Bendale and T. E. Boult. Towards open set deep networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1563–1572, 2016.