Class-Specific Semantic Reconstruction for Open Set Recognition

Hongzhi Huang, Yu Wang , Qinghua Hu , SeniorMember, IEEE, and Ming-Ming Cheng , SeniorMember, IEEE

摘要

开放集识别使深度神经网络 (Deep Neural Networks, DNN) 能够识别未知类别的样本，同时对已知类别的样本保持较高的分类精度。现有的基于自动编码器 (Auto-Encoder, AE) 和原型学习的方法在处理这一具有挑战性的任务方面显示了巨大的潜力。本文提出了一种新的方法，称为特定类语义重构 (Class-Specific Semantic Reconstruction, CSSR)，它集成了自编码器和原型学习的能力。具体来说，CSSR 用特定于类的自编码器表示的流形替换原型点。与传统的基于原型的方法不同，CSSR 对每个已知类进行建模，并通过重构误差来度量类的归属。将特定类别的自编码器插入到 DNN 主干的顶部，重建 DNN 学习到的语义表示，而不是原始图像。通过端到端学习，DNN 和自编码器相互促进，学习有区别的和有代表性的信息。在多数据集上进行的实验结果表明，该方法在封闭集和开放集识别方面都取得了良好的性能，并且具有足够的简单性和灵活性，可以整合到现有框架中。

关键词： 分类；开放集识别；自动编码器；原型学习；特定类语义重构

1 引言

传统的 DNN 是基于封闭集假设进行训练的，测试类都在训练过程中出现。在实际应用程序中，测试示例可能来自未知的类 [22]。当遇到这样一个未知样本时，传统的 DNN 会强制将其归类为已知类之一，并做出错误的预测，在某些关键场景下，如医疗诊断和自动驾驶，可能会造成无法弥补的损失。OSR 通过使模型能够正确地对已知类 (即封闭集) 的样本进行分类，并准确地识别未知类 (即开放集) 的样本来解决这一难题。

OSR 通过使模型能够正确地对已知类 (即封闭集) 的样本进行分类，并准确地识别未知类 (即开放集) 的样本 [5] 来解决这一难题。OSR 面临的主要挑战是，在训练过程中没有关于未知类的信息，这使得很难区分已知和未知类 (即，减少开放空间风险) [15]。传统的 DNN 强调已知类的判别特征，并学习整个特征空间的一个分区。这导致了一个严重的问题: 未知类的样本仍然位于特定的区域，因此，被识别为具有高置信度 [20] 的已知类。因此，之前的许多工作都提出了学习已知类的紧凑表示，从而使模型能够分离封闭集空间和开放集空间。其中，基于 AE 的方法 [12]、 [16]、 [21] 和类原型方法 [2]、 [3]、 [20] 是目前最强大的。

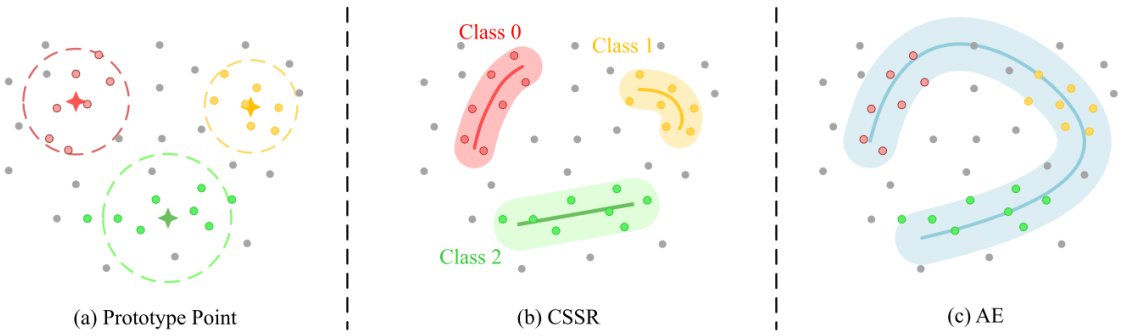


图 1. 不同方法对比

如图 1(c) 所示，基于 AE 的方法通过重建原始输入来学习潜在表示，以保留图像中的大部分信息。由于 AE 在训练时可以学习对已知类的图像进行重构，所以来自未知类的测试图像重构误差较大，因此可以识别出未知类。这可以被认为是学习一个低维流形来适应已知样本的分布。为了对已知类进行分类，这些方法基于原始图像像素级重建得到的潜在表示学习分类器。但这些方法存在两个问题:(1) 分类退化 (2) 开放空间风险引入。分类退化是指将 AE 学习到的潜在表示用于分类器会损害封闭集分类的性能，这主要是因为一些不必要的分类信息 (如背景信息) 被保留，干扰了分类器识别已知类的学习。开放空间风险引入是指为了拟合已知样本而学习的连续流形可能吞噬类间区域，从而引入开放空间风险。

与基于 AE 的方法不同，类原型方法，包括广义卷积原型学习 [20] 和最近提出的互反点学习 [3]，分别学习类特定的点来适应提取的对应于标记类或休息类的表示。这些方法的意义很简单。然而，原型学习框架在 OSR 任务中仍然面临着巨大的挑战。主要的挑战是类的不充分表示问题，在这种情况下，仅使用一个点或很少的点不能充分表示类。一方面，原型学习假设类特有特征为高斯分布。然而，这在现实世界的应用程序中很少得到满足，这将引入开放空间的风险。另一方面，在原型学习框架中，类内特性被压缩到相当有限的数量，这可能会导致模型过滤掉某些必要的帮助识别未知类的信息。

为了解决上述基于 AE 和类原型学习方法的问题，作者提出了一种新的方法，称为类特定语义重构 (CSSR)，图示描述如图 1(b) 所示。具体来说，CSSR 通过使用特定的 AE 流形对每个已知类进行建模，从而降低开放空间风险。在 CSSR 中，使用 DNN 提取每个样本的特征，然后，为每个已知的类指定一个单独的 AE，以将不同的类投影到不同的流形中。AE 被插入 DNN 骨干的顶部来重建语义表示，而不是原始图像。由于流形是特定类别的一种可学习表示，因此通过点到流形距离的重构误差归一化来进行分类。CSSR 使交叉熵损失最小化，使标记类别的声发射的重建误差最小。

作者提出的框架可以帮助解决现有方法的上述问题。与基于 AE 的方法相比，本文提出的 CSSR 方法通过丢弃不必要的信息并重构语义特征 (而不是原始图像) 来解决分类退化问题，通过学习类特定流形来释放被覆盖的类间区域来处理开放空间风险问题。与类原型方法相比，CSSR 方法通过学习类特定流形，很好地解决了类的欠表示问题。这不仅打破了类的高斯假设，而且保留了类的更多关键信息，而不是用一个点表示类。通过端到端学习过程，类特定的 AE 和 DNN 在学习高度类相关语义表示的同时，相互促进识别开放空间。AE 倾向于将每个类与语义特征的子集相关联，一个已知类的样本倾向于激活其相关的特性，而不激活不相关的特性。对于未知类的样本，它们的语义特征不会被激活，因为它们与已知类的任何特征都不相关。在不同数据集上进行的实验结果表明，该方法的识别性能明显优于其他先进

的方法，提高了封闭集和开放集的识别性能。

2 相关工作

论文的工作主要涉及开放集识别，特别是基于 AE 和类原型的方法。开放集识别自然与其他一些问题有关 (如异常检测和新颖点检测)。本节简要讨论了异常检测的检测方法。

2.1 开放集识别

早期的开放集识别工作使用传统的机器学习方法。通过测量样本和已知类别之间的相似性，使用分类器产生的分数来识别出未知样本。例如，Scheirer 等人 [15] 采用支持向量机对已知类进行识别，采用极值分布对未知类进行检测。近年来，DNN 强大的表示学习能力被应用于未知信息的检测。

已有学者设计或利用了 OSR 的分类层。一个简单的选择是利用最大 SoftMax 概率并拒绝不自信的预测 [8]。Bendale 等人 [1] 证明了 SoftMax 概率的不鲁棒性，提出用 OpenMax 函数代替 SoftMax 函数，OpenMax 函数重新分配 SoftMax 的分数，明确地得到未知类的置信分数。Zhou 等人 [23] 提出了占位符学习的概念，通过为未知类预留分类器占位符来校准过度自信的预测。

2.1.1 基于 AE 的 DNN 方法

Zhang et al. [22] 认为重构误差包含有用的判别信息，并提出使用稀疏表示对开放集识别问题进行建模。Yoshihashi 等人 [21] 设计了 CROSR 方法，该方法使用潜表示进行封闭集分类器训练和未知检测。Oza 和 Patel [12] 提出了一种两步 C2AE 方法。该方法首先对编码器进行封闭集识别训练，然后保持编码器固定，并添加类条件信息对解码器进行未知检测训练。Sun 等人 [16] 利用变分声发射迫使不同的潜在特征逼近不同的高斯模型进行未知检测。他们随后开发了 CPGM [17]，将判别信息添加到概率生成模型中。Perera 等人 [13] 将原始图像和重建图像送入一个分类网络，当重建图像与原始输入一致时，预测同时是可信的。然而，基于 AE 的方法存在两个问题：(1) 从像素级图像重建中学习到的表示包含不必要的背景信息，这可能会损害封闭集和开放集的性能。(2) AE 学习一个连续流形来拟合已知样本，这可能会吞噬类间区域。

2.1.2 类原型 DNN 方法

Yang 等人 [20] 提出了广义卷积原型学习，用面向开放世界的原型模型代替了封闭世界假设的 SoftMax 分类器。Chen et al. [3] 提出了倒易点学习 (RPL)，它根据倒易点的差异性将样本分类为已知或未知。随后，将 RPL 进一步改进为 ARPL [2]，加入额外的对抗训练策略，通过生成混淆训练样本，增强模型对已知和未知类的可分辨性。由于拟合能力和表示多样性的不足，类原型方法受到限制，作者使用特定类的 AE 来解决这个问题。

2.2 OOD 检测

OOD 检测首先由 Hendrycks 和 Gimpel [8] 提出，它涉及到对不属于训练集的样本进行检测。有几种方法考虑了在训练过程中 OOD 样本可用的问题，然而，这与我们的任务不一致，因为在训练期间只能访问分布中的数据。虽然利用已知的离群值可以极大地简化 OOD 检测，但开放集识别在实际应用中更为常见和现实。此外，设计良好的模型也可以从异常值暴露中受益，例如 OpenGAN [10] 可以使用或不使用辅助 OOD 数据。接下来，我们主要关注在没有额外 OOD 数据的情况下训练的模型。

2.2.1 有监督方法

与开放集识别类似的问题设置，这些方法在分类任务的基础上构建 OOD 检测器。一些方法寻求更好的评分函数，包括最大 SoftMax 概率 [8]、最大对数评分 [7] 和能量评分 [11]。Vyas 等人 [19] 使用一种遗漏分类器的集合来单独模拟 OOD 可访问训练。Sastry 和 Oore [14] 提出用 Gram 矩阵来表征活动模式，并通过计算训练数据中 Gram 矩阵的元素方向偏差来评分。

2.2.2 自监督方法

这些方法都是比较新颖的，它们通过半监督来利用已经学得很好的表示。Golan and El-Yaniv [6] 和 Hendrycks 等人 [9] 考虑的是预测图像变换的任务 (例如，将图像旋转)，这在 [18] 中也被用作辅助任务。自监督对比学习在无监督表示学习 [4] 中取得了相当大的成功，并被应用于 OOD 检测。他们观察到通过对比学习获得的表征在内分布和外分布数据之间有明显的模式。虽然这些方法是为无标记设置设计的，但也被扩展为监督学习。

3 本文方法

3.1 本文方法概述

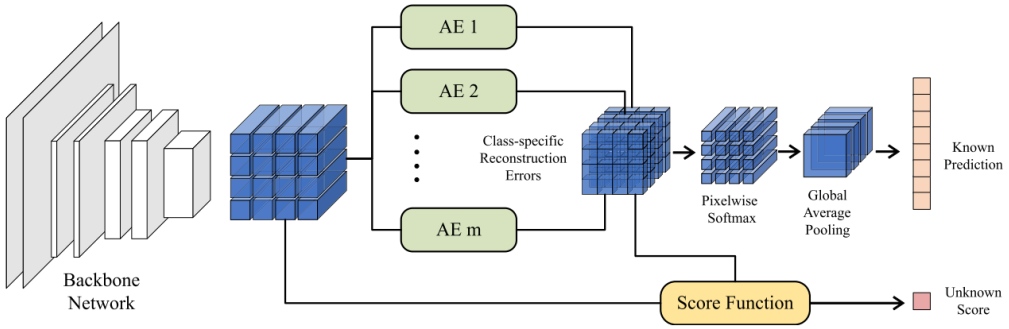


Fig. 5. Overall architecture of our proposed model. The backbone network ($\mathcal{B}$) takes an image as input and extracts its semantic feature map $\mathbf{z}$. The AE ($\mathcal{A}_c$) with respect to class c encodes and reconstructs the pixelwise semantic representation $\mathbf{z}$. Subsequently, we take the pixelwise reconstruction errors of class-specific AEs as logits and make pixelwise SoftMax on the logits multiplied by γ. Then, global average pooling is applied to reduce the pixelwise predictions to a general prediction for the whole image. For *unknown inference*, the model uses the pixelwise reconstruction errors corresponding to the predicted class and semantic features as input, and scores the unknownness for the image. Finally, a threshold is determined ensuring 95% known samples are correctly accepted; samples are rejected if their unknown scores are below the threshold.

图 2. CSSR 总体框架

如图 2 显示了 CSSR 的训练和推理过程。CSSR 框架包括两个主要模块: 用于学习潜在表示的骨干网 B 和用于分类已知类和检测未知类的特定类的 $AEs = \{\mathcal{A}_i\}_{i=1}^{m}$。

为了充分利用从 B 中提取的语义特征图 $Z = \{\mathbf{z}_{ij}\}$，平等地对待每个像素的潜在表示，然后通过平均像素预测来进行全局预测:

$$p(y = i|Z, \mathcal{A}) = \frac{1}{|Z|} \sum_{\mathbf{z} \in Z} p(y = i|\mathbf{z}, \mathcal{A}). \tag{1}$$

显然，$p(y = i|Z, \mathcal{A})$ 之和等于 1，并且 $p(y = i|z, \mathcal{A})$ 之和分别等于一个。最后，通过梯度下降最小化真类 c 的负对数概率来训练模型，训练方法如下:

$$\mathcal{L} = -\log p(y = c|Z, \mathcal{A}). \tag{2}$$

由于每个像素的表示都集中在输入图像的一个局部区域上，这些操作可以看作是对原始输入进行软剪切和增强后的图像集合预测结果的增强。在测试阶段，它自然地应用了测试增强技术，从而提高了性能。上述操作可以通过 1*1 卷积、像素化 SoftMax 和全局平均池来实现。此外，为了简单起见，采用线性编码器和解码器来实现 AEs。

该方法还可以扩展到一个倒易学习框架。特定于类的 AEs 可以用来代替倒易点集来估计其差异性。唯一的修改是设置一个负超参数 g，然后，假设 $d(\mathbf{z}, \mathcal{A}_c)$ 最大时，重构误差可近似为

$$d(\mathbf{z}, \mathcal{A}_c) = \|\mathbf{z} - \mathcal{A}_c(\mathbf{z})\|_1 \approx \max_{\mathbf{v} \in V_c} \|\mathbf{z} - \mathbf{v}\|_1. \tag{3}$$

在本文的其余部分，将 CSSR 的互易版本称为 RCSSR。

3.2 类特定语义重构

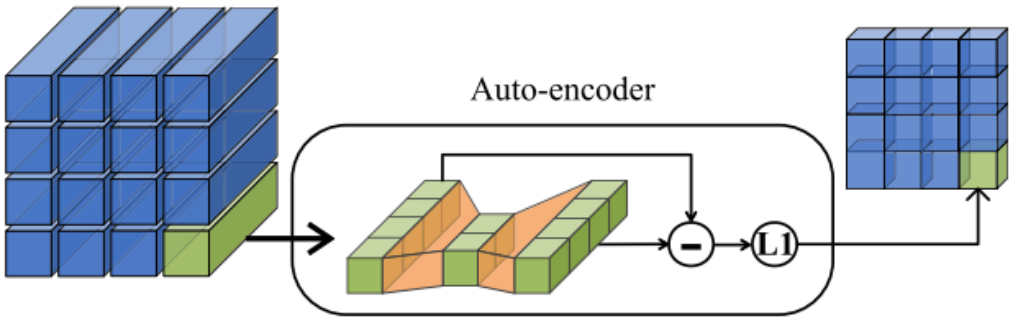


Fig. 2. Structure of individual AEs. Note that we operate the reconstruction process at the pixel level. It takes the semantic feature map as input and outputs the reconstruction error at each pixel.

图 3. 单个 AE 结构

对于每个类别 i(表示为 A_i)，我们用 AE 替换特定于类的点集 U_i。如图 3 所示，以潜表示 z 为输入，输出重构表征 $\hat{\mathbf{z}} = \mathcal{A}_i(\mathbf{z})$。然后，用 L_1 范数计算重构误差为

$$d(\mathbf{z}, \mathcal{A}_i) = \|\mathbf{z} - \mathcal{A}_i(\mathbf{z})\|_1. \tag{4}$$

在原型学习的基础上,基于重构误差,我们的框架可以估计类的归属。给予样本 $(\mathbf{x}, c) \in \mathcal{X}$,设 $p(y = i|\mathbf{x}) \propto (-d(\mathbf{z}, \mathcal{A}_i))$ 学习流形原型。通过应用 SoftMax 对对数进行归一化，最终概率可以定义为

$$p(y = i|\mathbf{z}, \mathcal{A}) = \frac{e^{-\gamma d(\mathbf{z}, \mathcal{A}_i)}}{\sum_{j=1}^{m} e^{-\gamma d(\mathbf{z}, \mathcal{A}_j)}}, \tag{5}$$

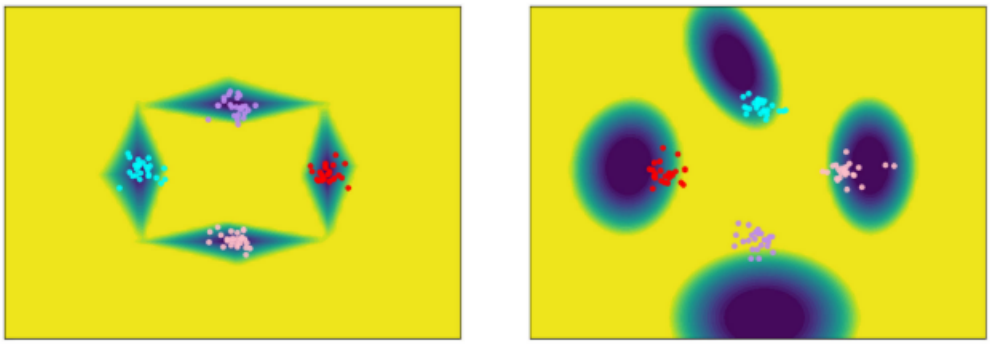
其中 γ 是控制概率分配的硬超参数。考虑到一个理想的解决方案，最大限度地提高地真类别的输出概率，AE 应该首先学习最小距离映射 (从特征到流形)，以最小化地真类别的重建误差。同时，多组 AE 也应该学会彼此保持距离，以最大限度地提高除与真实标签相对应的 AEs 外的重构误差。

类特定的 AEA_i 定义了类特定的流形 V_i。在上述理想情况下，最大化 $p(y = c|\mathbf{x}, \mathcal{A})$ 可以认为是具有无限个原型点 (流形 V_i) 的原型学习。$d(\mathbf{z}, \mathcal{A}_c)$ 最小，重构误差可近似表示为

$$d(\mathbf{z}, \mathcal{A}_c) = \|\mathbf{z} - \mathcal{A}_c(\mathbf{z})\|_1 \approx \min_{\mathbf{v} \in V_c} \|\mathbf{z} - \mathbf{v}\|_1. \tag{6}$$

3.3 管理开放空间风险

3.3.1 拟合已知类



(a) CSSR (b) CSSR MSE

Fig. 3. Comparison of open space modeled by (a) CSSR, and (b) CSSR with mean squared reconstruction error. We set up a trivial experiment, where four classes generated from different Gaussian distributions are used for training. Then, we tested the entire feature space for unknown detection. The bright yellow and dark blue regions correspond to the open space and the close space identified by the different methods, respectively.

图 4. (a) CSSR 模拟的开放空间和 (b) CSSR 均方重建误差的比较

在原型学习中，使用均方误差 (MSE) 作为距离度量。在使用 MSE 时，原型点和语义特征之间的不一致分布很可能被学习到。图 4(b) 中的例子将这种现象可视化，原型与真实标签

分布之间出现了差距。我们还观察到，这种差距是由于模型的过拟合，而不是欠拟合。接下来分析 MSE 是如何导致上述不一致的，而平均绝对误差保持了一致性。

在接下来的分析中，考虑最简单的原型学习形式，其中 $|U_i| = 1$ 和 $\mathbf{u}_i \in U_i$ 代表了类 i 的唯一原型点，因此对 $d(\mathbf{z}, U_i)$ 简化为 $\|\mathbf{z} - \mathbf{u}_i\|$，并且第 i 类的 SoftMax 概率为

$$p(y = i|\mathbf{z}, U) = \frac{\exp(-\|\mathbf{z} - \mathbf{u}_i\|)}{\sum_j \exp(-\|\mathbf{z} - \mathbf{u}_j\|)}. \tag{7}$$

原型学习的目的是，如果样本正好位于原型点 $\mathbf{z} = \mathbf{u}_c$，则使 $p(y = c|\mathbf{z}, U)$ 最大化。当偏移量 $\varepsilon = \mathbf{z} - \mathbf{u}_c$ 变大时，概率应该降低。然而，下面的定理表明，使用均方误差是无法达到上述目的的。

Theorem 1. *$d(\mathbf{z}, U_i) = \|\mathbf{z}-\mathbf{u}_i\|_2^2$, assuming $\mathbf{u}_i \neq \mathbf{u}_j$ for $\forall i \neq j$, there exists c and $\varepsilon \neq 0$ satisfying $p(y{=}c|u_c, U) < p(y = c|\mathbf{u}_c + \varepsilon, U)$.*

Theorem 2. *$d(\mathbf{z}, U_i) = \|\mathbf{z} - \mathbf{u}_i\|_1$, for each c,ε,$p(y = c|\mathbf{u}_c, U) \geq p(y = c|\mathbf{u}_c + \varepsilon, U)$ stands.*

定理 1 表明，使用 MSE 时，$\mathbf{z} = \mathbf{u}_c$ 可能不是最大化 $p(y = c|\mathbf{z}, U)$ 并且使交叉熵损失最小化的最佳解，从而导致上述分布不一致。如定理 2 所示，$p(y = c|\mathbf{u}_c, U) \geq p(y = c|\mathbf{u}_c + \varepsilon, U)$，因此在 $\mathbf{z} = \mathbf{u}_c$ 时取最大的 $p(y = c|\mathbf{z}, U)$，最小的交叉熵损失，保证了语义特征分布与原型点的一致性。图 4(a) 所示的实验表明，原型拟合良好，开放空间风险得到了很好的缓和。

3.3.2 学习类相关的特征

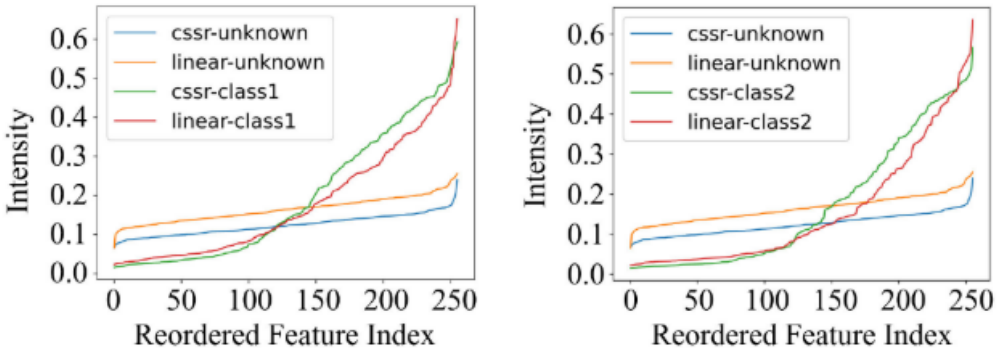


Fig. 4. Visualization of the class-specific feature activation for CSSR. We took six categories from Cifar10 as known classes and the remaining four as unknown. Note that the magnitude for each feature is normalized across the six known classes; we sorted each curve to have increasing order for better visualization.

图 5. CSSR 激活特定类特征的可视化

在骨干网和类特定 AE 的联合优化过程中，学习过程中双方都进行优化。在训练 AE 拟合已知类的同时，对骨干网进行训练，使提取的特征接近类特定流形。这种多样性自然地表现在位于流形表面的特征上。同时在流形的垂直方向上，重构误差使得表示法更加紧凑。这

些属性使语义特征与类相关。每个类别都由一个全局特征子集描述并与之关联，一个示例倾向于只激活与其类别对应的相关特征。

在这里给出一个直观的例子来解释特定类的特性如何更容易学习。考虑这样一种情况，AE 具有最简单的形式: 每个编码器都是特定于类的特性子集的标识映射，解码器以相同的方式映射回这些特性。激活类相关的特性不会导致重构错误，而激活类不相关的特性会导致重构错误。为了减少这种情况下的重构误差，主干学习只激活类相关的特征。对于联合优化，骨干网的拟合能力强得多，可以降低拟合复杂度。最后，模型最终提取类特定的特征，这些特征也可被用来检测未知类。

图 5 显示了已知类和未知类的平均激活强度 (即绝对激活值)。已知的类在某些特定特性上被强烈激活，而在其他特性上几乎没有被激活。然而，未知类在所有特性上的激活值都很低，因为它们没有经过训练，也没有与任何特定的特性相关联。与普通线性分类层相比，CSSR 提取的特征在以下几个方面具有优势:(1)CSSR 的类相关特征贡献更统一，而不是集中在几个特征上;(2) CSSR 的类无关特征激活较少;(3) CSSR 的未知类激活强度明显低于普通线性分类器，表明已知类与学习到的语义特征之间的关联更强。

3.4 未知类检测

假设给出了测试样本的语义特征 Z 和预测标签 c。从两个不同的角度构造了用于未知检测的评分函数:(1) 重构误差和 (2) 类特有特征统计量。

3.4.1 基于重构误差的评分函数

一个自然的想法是利用重构错误来检测未知的类。如 3.2.2 节所述，未知样本会导致语义特征失效。缺乏激活的语义特征导致低重构错误，从而检测未知类失败。在使用线性解码器和线性编码器实现 As 时，考虑了 $\|\mathbf{z}\|_1$ 和 $\|\mathbf{z}-\mathcal{A}_c(\mathbf{z})\|_1$ 之间的近似线性关系。编码器 f 和解码器 g 满足 $f(\lambda\mathbf{z})=\lambda f(\mathbf{z})$;$g(\lambda\mathbf{z})=\lambda g(\mathbf{z})$，因为它们是线性函数; 对于 tanh 激活，假设 $f(z)$ 位于 0 附近，$\tanh(\lambda\mathbf{z})\approx\lambda\tanh(\mathbf{z})$。因此，

$$\|\lambda\mathbf{z}-\mathcal{A}_c(\lambda\mathbf{z})\|_1\approx\lambda\|\mathbf{z}-\mathcal{A}_c(\mathbf{z})\|_1, \tag{8}$$

通过除以 $\|\mathbf{z}\|_1$ 来去除比例因子 λ。对于未知样本，z 位于零附近的要求很容易满足。此外，我们通过乘以一个额外的 $\|\mathbf{z}\|_1$ 项来考虑特征大小的未知检测能力。

具体来说，对于 CSSR，已知类应具有较低的相对重构误差，同时具有较高的特征量，定义第一个分数函数如下所示:

$$s_{p1}(\mathbf{z},c)=-\frac{d(\mathbf{z},\mathcal{A}_c)}{\|\mathbf{z}\|_1^2}. \tag{9}$$

同时对于 RCSSR，已知类样本应具有较高的相对重构误差和较高的特征幅值:

$$s_{r1}(\mathbf{z},c)=\frac{d(\mathbf{z},\mathcal{A}_c)}{\|\mathbf{z}\|_1}\times\|\mathbf{z}\|_1=d(\mathbf{z},\mathcal{A}_c). \tag{10}$$

与闭集分类过程类似，充分利用像素特征，对特征图 Z 像素化评分，将单个评分相加为整幅图像的最终评分，即 $s_*(Z,c)=\frac{1}{|Z|}\sum_{\mathbf{z}\in Z}s_*(\mathbf{z},c)$，其中其中 s_* 表示 s_{p1} 或 s_{r1}

3.4.2 基于激活模式的评分函数

虽然 CSSR 学习与类相关的特征的性质通过考虑特征的大小已经被利用了，通过考虑一阶和二阶统计量提出了一个更精细的模型的类特定的激活模式。低阶统计量已被应用于 OOD 检测中。而统计信息作为 OOD 分类器的输入特征，需要 OOD 样本进行训练。相反，我们收集统计信息，直接制定得分函数。

假设由训练集中的样本 x_i 获得语义特征映射 Z_i 和预测类 c_i。由于只关心特征的激活强度，所以通过计算特征元素的绝对值来对所有特征图进行预处理; 在接下来的讨论中，假设特征图已经进行了预处理。因为需要特定于类的模式，所以特征映射根据其预测的类被分组到不同的集合中，即特征集合 $\mathcal{Z}^c = \{Z_i | c_i = c, i = 1, 2, ..., n\}, c = 1, ..., m$。

对于一阶统计量，我们首先取类特有的平均激活强度:

$$\mu_i = \sum_{Z \in \mathcal{Z}^i} \sum_{\mathbf{z} \in Z} \frac{1}{|\mathcal{Z}^i||Z|} \mathbf{z}. \tag{11}$$

为了考虑不同特征激活强度的不同尺度，进一步采用归一化交叉分类:

$$\tilde{\mu}_i = \frac{\mu_i}{\sum_j \mu_j}, \tag{12}$$

其中向量除法是按元素进行的。为了检测未知，激活一个已知类所激活的样本更有可能是同一个已知类。其中，每个特征的激活强度用 $\tilde{\mu}_c$ 加权，未知量用各特征加权平均强度来计算。同时也执行像素级的分数整合。分数函数被正式定义为:

$$s_2(Z, c) = \sum_{\mathbf{z} \in Z} \frac{1}{|Z|} \mathbf{z}^\top \tilde{\mu}_c. \tag{13}$$

对于二阶统计量，利用 Gram 矩阵模型间特征共现。设 $F \in \mathbb{R}^{D \times |Z|}$ 为特征图 Z 中像素向量串列的特征强度矩阵，D 为特征维数。第 i 个样本的 Gram 矩阵由 $G = FF^\top$ 定义。Gram 矩阵 G 中的元素描述了相应的两个特征 (按行和列索引) 可能同时被激活的程度。我们平均特征映射类的 Gram 矩阵——特别是作为特征共现模式的模板。然后，对于一个测试样本，我们计算其 Gram 矩阵，并使用预先计算的模板对其预测标签进行元素乘法的和来对其未知度进行评分。该过程可以表述为

$$G^c = \frac{1}{|\mathcal{Z}^c|} \sum_{Z \in \mathcal{Z}^c} G(Z), \tag{14}$$

$$s_3(Z, c) = Sum(G^c \odot G(Z)), \tag{15}$$

其中 $Sum(\cdot)$ 是一个对矩阵元素求和的函数，是矩阵元素乘法。观察到上述操作等价于将像素特征 z 扩展到一个二阶多项式空间，即 $\mathbf{z}\mathbf{z}^\top = [z_i z_j]_{D \times D}$ 。Gram 矩阵可以写成像素扩展特征 $G = \sum_{\mathbf{z} \in Z} \mathbf{z}\mathbf{z}^\top$ 的和; 因此，G^c 表示扩展特性空间的一阶统计信息。此外，此外，评分函数可以看作像素评分和积分 $s_3(Z, c) = \sum_{\mathbf{z} \in Z} Sum(G^c \odot \mathbf{z}\mathbf{z}^\top)$。除了 Gram 矩阵的基本定义外，提出的扩展的高阶 Gram 矩阵在这里是可选的，即:

$$G = \left(F^p F^{p\top}\right)^{\frac{1}{p}}, \tag{16}$$

3.4.3 完整评分函数

在对以上分数进行整合之前，我们对单个分数进行归一化处理，从而统一不同分数函数的尺度。具体来说，对于每个评分函数 s_*，我们对随机增强训练样本进行评分，以减少过度自信对拟合良好的训练样本的影响。然后，计算平均值和标准差，并将其归一化。该过程可以表示为

$$\tilde{s}_*(Z,c) = \frac{s_*(Z,c) - E(s_*)}{Std(s_*)}, \tag{17}$$

上式中，$E(s_*)$ 和 $Std(s_*)$ 分别表示对 s_* 的预计算平均值和标准差。我们现在通过线性组合来整合这三个分数函数，得到最终的分数函数:

$$s_{all}(Z,c) = w_1 \times \tilde{s}_{*1}(Z,c) + w_2 \times \tilde{s}_2(Z,c) + w_3 \times \tilde{s}_3(Z,c), \tag{18}$$

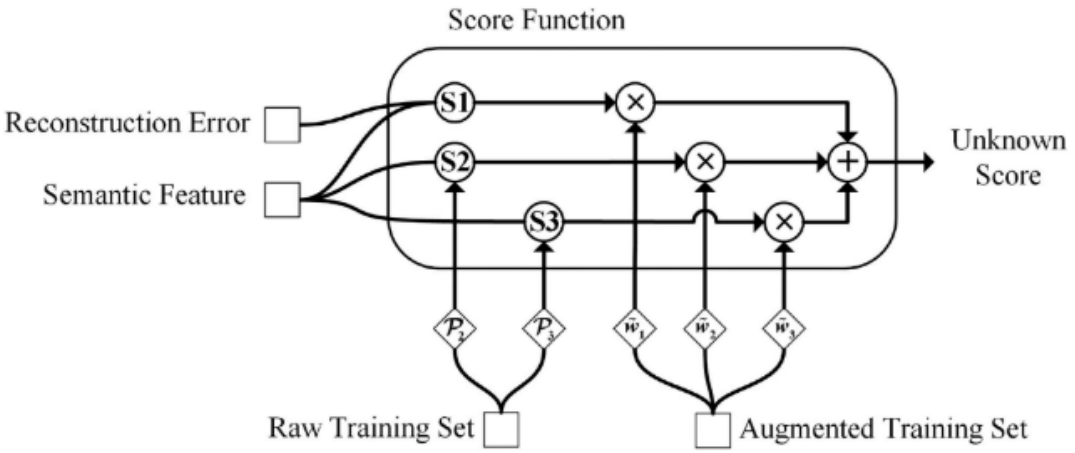


Fig. 6. Unknown inference process. Before inference on test samples, we obtain first- and second-order statistics (corresponding to P2 and P3 in the figure, respectively) from the raw training set. Then, the normalization coefficients ($\frac{1}{Std(s_*)}$) are obtained from the augmented training set for individual scores. The final score weights are calculated by the product of normalization coefficients and predefined weights, i.e., $\tilde{w}_* = \frac{w_*}{Std(s_*)}$, which forms the score function by $s_{all}(Z,c) = \sum_i \tilde{w}_i \times s_i(Z,c)$.

图 6. 未知分数推理过程

过程如图 6 所示。将未知检测的过程总结为:(1) 在不增加数据的情况下遍历训练集，收集语义特征的一阶和二阶统计量。这个步骤使用非扩充训练样本保存已知类的完整信息。(2) 对训练集进行数据增广，收集单个分数的归一化参数。该步骤使用增强训练样本来减少过度自信的影响。(3) 利用公式 (18) 进行推理。为了拒绝未知类样本，取一个阈值，保证 95% 的已知样本被接受。

4 复现细节

4.1 与已有开源代码对比

利用现有的开源代码，进行复现，开源代码来自：https://github.com/xyzedd/CSSR

5 实验结果分析

表 1. 复现对比

		AUROC		CLOSE ACC	
		CIFAR10	SVHN	CIFAR10	SVHN
Paper	CSSR	91.3	97.9	96.8	99.1
	RCSSR	91.5	97.8	97.0	99.1
Ours	CSSR	89.6	97.6	97.0	98.2
	RCSSR	91.8	97.6	97.1	98.2

复现结果如表 1 所示，复现结果与原文相差不大。

6 总结与展望

文章通过结合 AE 和原型学习框架，提出了一种新的端对端学习深度网络 CSSR，用与开放集识别任务。CSSR 为每个已知的类指定一个单独的 AE，作为传统原型学习框架中类特定点集的替代品。在骨干网的顶端插入 AE，重构学习后的图像语义表示。同时注意到，通常用于原型学习的 MSE 距离可能会导致原型点与地面真实数据分布之间的不一致分布。为了解决这一问题，采用 MAE 距离代替均方差，保证原型点与数据点之间的一致性。

文章也提到类间 AE 的特征激活需要保持距离，而文章工作在对于这一点没有进行约束，下一步可以从此出发，对不同类的特征激活添加约束，使得类间特征激活保持距离。

参考文献

[1] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[2] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2022.

[3] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 507–522, Cham, 2020. Springer International Publishing.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[5] Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3614–3631, 2021.

[6] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[7] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings, 2022.

[8] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2018.

[9] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[10] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 813–822, October 2021.

[11] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020.

[12] Poojan Oza and Vishal M. Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Pramuditha Perera, Vlad I. Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M. Patel. Generative-discriminative feature representations for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[14] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8491–8501. PMLR, 13–18 Jul 2020.

[15] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[16] Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] Xin Sun, Chi Zhang, Guosheng Lin, and Keck-Voon Ling. Open set recognition with conditional probabilistic generative models, 2021.

[18] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11839–11852. Curran Associates, Inc., 2020.

[19] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L. Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[20] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, Qing Yang, and Cheng-Lin Liu. Convolutional prototype network for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2358–2370, 2022.

[21] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[22] He Zhang and Vishal M. Patel. Sparse representation-based open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1690–1696, 2017.

[23] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, June 2021.