arXiv:2303.02400v1 [cs.CV] 4 Mar 2023

真实场景中的细粒度 ImageNet 分类

玛丽亚 利姆佩赖乌[1], 康斯坦丁诺斯 托马斯[1] 和 乔戈斯斯塔穆[1]

[1]*AILS* 实验室，雅典国立理工大学电气与计算机工程学院

摘要

图像分类一直是深度学习中最热门的任务之一，每年都有大量令人瞩目的实现。然而，推动复杂架构不断提升性能指标的做法一直备受批评。鲁棒性测试可以揭示在常规模型评估阶段未被察觉的诸多漏洞和偏差。迄今为止，分布偏移下的模型鲁棒性主要是在精心筛选的数据集上进行考察。尽管如此，这类方法并未检验分类器在真实场景中的实际响应，例如提供未经筛选的网络爬取图像数据，其中包含相应类别的图像。在我们的工作中，我们借助层次知识识别密切相关的类别，并对其进行细粒度分类。我们在多种卷积和基于 Transformer 的架构上进行了广泛实验，揭示了模型在这一新颖设置下的鲁棒性。最后，我们再次运用层次知识来评估并解释误分类，从而提供一种信息丰富且适用于任何分类器的评估方案。

关键词

图像分类、知识图谱、鲁棒性、可解释评估

1. 引言

ImageNet [1] 一直是文献中最受欢迎的图像分类数据集之一，十多年来一直推动着各种流行模型实现的发展。ImageNet 分类领域的首次重大突破来自 AlexNet [2], ，这是一种用于图像分类的卷积神经网络（CNN），其性能大幅超越了竞争对手。此后，各种基于 CNN 的实现不断将准确率分数推向更高水平 [3]。

卷积滤波器的局部特性使其无法捕捉长距离视觉依赖关系，这被认为阻碍了性能的进一步提升，因此需要探索替代性的架构选择。为此，在自然语言处理中已成功应用的注意力机制 [4] 似乎成为卷积的有力替代方案，因为它们能够检测空间上相距较远的概念，并为其分配适当的重要性权重。事实上，对 Transformer [的改造

4] 面向视觉任务的应用，促成了视觉 Transformer（ViT）的问世 [5], 。它将图像划分为视觉图像块，并以类似原始 Transformer 处理词语的方式对其进行处理。因此，基于 Transformer 的图像分类器应运而生 [6, 7, 8, 9],，并取得了前所未有的最先进成果。

载于 *A. Martin*、*K. Hinkelmann*、*H.-G.Fill*、*A.Gerber*、*D. Lenat*、*R. Stolle*、*F. van Harmelen*（编），《*AAAI 2023* 春季研讨会论文集：需要结合机器学习与知识工程的挑战》（*AAAI-MAKE 2023*），旧金山机场凯悦酒店，加利福尼亚州，美国，*2023*年*3*月*27*日至*29*日。
marialymp@islab.ntua.gr（M. Lymperaiou）；kthomas@islab.ntua.gr （K. Thomas）；gstam@cs.ntua.gr （G. Stamou）0000-0001-9442-4186 （M. Lymperaiou）；0000-0002-7489-7776 （K. Thomas）；0000-0003-1210-9874 （G. Stamou）
©2023 年本文作者版权所有。根据知识共享署名 4.0 国际许可协议（CC BY 4.0），允许使用。CEUR 研讨会论文集（CEUR-WS.org）

尽管人们投入大量精力，通过采用日益精细的架构和技术来不断提升模型性能，这也不可避免地增加了训练所需的计算资源，但关于此类模型能否妥善应对分布偏移，仍有一些悬而未决的问题。分布偏移是指在与模型训练所用数据分布不同的分布上测试已训练好的模型。近年来，对分布偏移的分析日益受到关注 [10, 11, 12, 13, 14], ，因为这是提升模型鲁棒性的关键一步。大多数此类研究都采用像素级扰动，人为影响所研究的分布。然而，对人工分布偏移的设置限制过多，无法涵盖各种真实场景，阻碍了图像分类器实现鲁棒泛化。在这种情况下，自然偏移 [15, 16, 17, 18]更具代表性。它们通常需要创建一个经过精心筛选的数据集，其中包含视角或物体背景变化、旋转及其他细微变化等图像变体。合成偏移和自然偏移都可以纳入数据增强技术；在训练过程中使用这些技术 [19, 20, 21, 22]有助于开发鲁棒模型。

迄今为止，尚无方法测试“真实环境下”的图像分类，即面对与预定义数据集标签相对应的未经筛选的图像时进行分类。我们认为，这是一个面向现实世界用户的场景：需要对全新的、对应 ImageNet 标签的图像进行恰当分类。例如，网上找到的一张猫的图像可能与 ImageNet 中的猫图像实例存在显著差异，即使考虑到常见的分布偏移也是如此。尽管人类能够以令人满意的信心识别图像中的猫，我们仍不确定图像分类器能否做到这一点；未经筛选的图像可能存在的变化范围不受限制，因此需要具备先进的泛化能力，才能正确理解 ImageNet 类别真正的判别特征，而不被无关特征干扰。

当需要进行细粒度分类时，“真实环境下”的分类问题会变得更加困难，因为区分密切相关的类别依赖于细致的判别特征，而这些特征在未经筛选的场景中可能不那么常见。例如，暹罗猫和波斯猫在外观上有许多相似之处，这增加了学习并重现数据集偏差的风险，尤其是在存在分布偏移时。我们认为，这种风险源于现有分类器缺乏外部或领域知识，而这些知识可以帮助人类区分密切相关的类别。

总而言之，在本文中，我们希望回答以下问题：

1. 在 ImageNet 或网络图像上预训练的不同模型，在从 Google图片抓取的未经筛选图像集上（将 ImageNet 标签作为谷歌查询词）表现如何？我们通过构建一种基于未经筛选的网络图像的新型自然分布偏移来研究这一问题，并在其上评估各种图像分类器。2. 鉴于多个 ImageNet 类别之间存在层次关系，层次知识如何有助于评估分类结果？我们尝试验证这样一种假设在何种程度上成立：缺乏外部知识会限制分类器的泛化能力。因此，我们利用 WordNet [23] 发现给定术语的邻近词，并测试分类器是否难以区分密切相关的类别。3. 分类评估能否可解释？诸如 WordNet 等知识来源可以揭示概念（ImageNet 类别）之间的语义关系，从而提供

连接经常被混淆的类别的可能路径。

- 我们的代码可在 https://github.com/marialymperaiou/classification-in-the-wild.

2. 相关工作

图像分类器 随着用于分类任务的神经网络架构蓬勃发展，计算机视觉成为从近期发展中获益最多的领域之一。卷积分类器（卷积神经网络）是一种成熟的骨干网络，早期的成功尝试 [2] 已经为更精细的架构铺平道路，例如 VGG [24], Inception [25], ResNet [26],Xception [27], InceptionResnet [28] 等 [3]。尽管 ConvNeXt[29] 等当代研究重新审视并坚持经典范式，取得了先进性能，但围绕使用卷积神经网络进行图像分类仍存在一些批评。Transformer框架 [4]通过采用自注意力机制，在自然语言处理应用中广泛取代了先前的架构；这一快速进展启发研究者将类似模型用于计算机视觉，以回应上述批评 [30]。因此，基于 [4] 构建的视觉Transformer（ViT） [5] 为相关研究确立了新的基准；此后，涌现出多种相关架构。总体而言，基于 Transformer 的模型依赖大量训练数据来确保良好的泛化能力。DeiT [31], 放宽了这一要求，使模型能够在中等规模的数据集上进行学习。后续研究引入了新颖的基于 Transformer 的架构，例如 BeiT [9], Swin [32] 和 RegNet [33], ，通过特定改进提升性能。总体而言，已有证据表明，与经典的卷积神经网络图像分类器相比，视觉Transformer具有更强的鲁棒性 [34]。在我们的工作中，我们通过在未经筛选的细粒度设置下测试卷积神经网络和基于 Transformer 的分类器，验证这一论断在多大程度上成立。

分布偏移下的鲁棒性现有图像分类器的泛化能力一直是一个关键问题， [35], 目前已有一些研究从不同视角对此展开探讨。针对精心整理的数据进行人工损坏 [36, 14, 37, 16, 11] 或自然偏移 [15, 38] ，已经揭示了偏差和架构漏洞。对抗鲁棒性 [39, 40, 41, 42, 43]是一个相关领域，研究者会测试模型抵御对抗样本的能力；这类样本会对图像施加难以察觉却影响显著的扰动。与这些研究不同，我们聚焦于未经筛选的图像数据所引发的自然分布偏移。在架构选择方面，许多研究开展鲁棒性测试，试图解决卷积神经网络与 Transformer 之争 [34, 44, 45], ；另一些研究则着重于解释和理解模型鲁棒性 [46, 47, 48]。我们的方法通过对卷积神经网络和基于 Transformer 的架构进行实验，将这类研究思路应用于未经筛选的数据场景。

3. 方法

我们方法的整体流程（图 1）分为三个阶段。首先，应收集常见术语（查询词）及其作为 ImageNet 类别存在的子类别，构建数据集。然后，从 Google搜索中爬取与这些术语对应的图像。在第二阶段，利用各种预训练分类器对爬取图像进行分类。

报告给定类别之间的层级关系，以丰富评估过程。最后，收集误分类样本之间的所有语义关系，从而提取解释，并量化错误预测的类别与其真实标签之间的偏差。

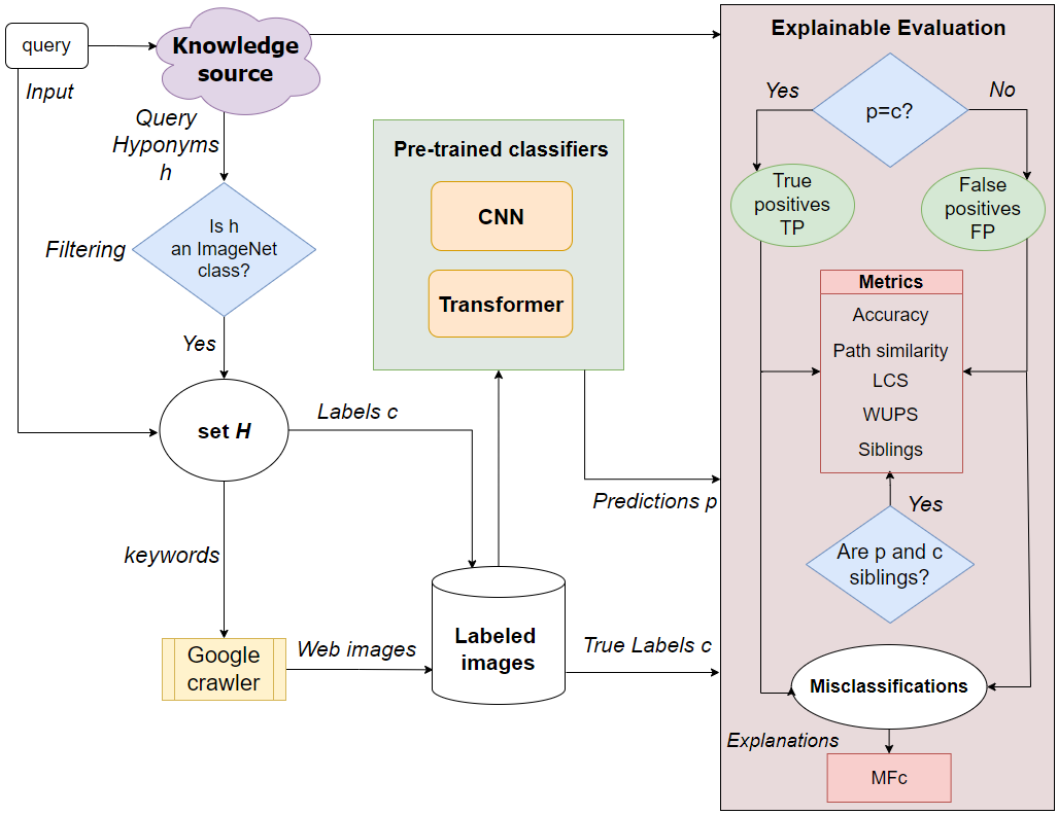

图1：我们方法的概述。

数据集创建 我们首先收集用户定义的、与视觉概念相关的常用词，将其作为查询词，并以此为起点提取子类别。我们利用 WordNet 层级结构 [23]，通过上位词—下位词（IsA）关系获取子类别；这两者分别指代更一般或更具体的概念。例如，给定查询词“汽车”，其上位词是“机动车”（“汽车” IsA “机动车”），而其下位词包括“豪华轿车”（“豪华轿车” IsA “汽车”）、“跑车”（“跑车” IsA “汽车”）以及其他具体车型。因此，我们将查询词映射到 WordNet，以获取它们所有的直接下位词，从而构建下位词集合 H。随后，我们筛除不属于 ImageNet类别标签的下位词。

将初始查询中经过筛选的 H 类别作为搜索词，提供给适用于搜索 Google图片的网络爬虫。我们为返回的 Google图片数量设置了预定义阈值 k，以便在包含几乎相同数量样本的类别上评估分类器。这是必要的，因为某些热门类别返回的

Google图片可能比其他类别多得多。我们将尝试*k*的多个取值，从而影响关键词相关性与数据集规模是否充足之间的权衡。检索到的图像构成一个带标签的数据集 D，关键词即为标签。

分类 我们考察了多种图像分类器，以测试它们在未经筛选的网络图像上进行细粒度分类的能力。我们首先以基于卷积的模型作为基线模型开展实验，因为人们普遍认为，这类模型对分布偏移和其他扰动的鲁棒性较差 [34],，随后再考察近期的基于 Transformer 的架构。我们没有对所选模型进行进一步训练或微调。

对于每个模型，我们都在构成数据集的爬取图像上进行推理，具体如上一段所述。我们采用丰富的评估方案，以深入了解分类过程的各个方面。准确率可作为基准指标，帮助我们将研究结果与预期的分类结果进行比较。WordNet 相似度函数有助于分析误分类；例如，假设某个样本的真实标签是“猫”，而分类器在一种情况下预测为“狗”，在另一种情况下预测为“飞机”。直观而言，我们认为“猫”与“狗”的关系比与“飞机”更密切，因为二者都是动物。WordNet 层级结构体现了这种人类直觉，因此会根据概念在层级结构中的相关性赋予不同的惩罚。

这种基于概念的评估可以通过以下 WordNet 函数实现：路径相似度、Leacock-Chodorow 相似度（LCS）和 Wu-Palmer 相似度（WUPS）。**路径相似度**根据 WordNet 层级结构中连接两个概念的最短路径，评估这两个概念的相似程度。其取值范围为 0 到 1，1 表示可能达到的最高相似度分数。**LCS** 也用于寻找两个概念之间的最短路径，但还会考虑分类体系的深度。具体而言，公式1从数学上描述了两个概念 c_1 和 c_2之间的 LCS：

$$LCS = -\log \frac{path(c_1, c_2)}{2 \cdot d} \tag{1}$$

其中，$path(c_1, c_2)$ 表示连接 c_1 和 c_2 的最短路径，d 表示分类体系深度。LCS 值越高，表示概念之间的相似度越高。**WUPS** 会考虑两个概念 c_1 和 c_2 在 WordNet 分类体系中所处的深度，以及它们最具体的共同祖先节点的深度，该节点称为最近共同包含者。WUPS 分数越高，表示概念越相似。对于路径相似度、LCS 和 WUPS 这三项指标，我们分别计算其在所构建数据集 D的所有样本上的平均值。

此外，我们还报告误分类中兄弟概念所占的百分比。如果两个概念共享一个直接（1 跳）的父节点，则视为兄弟概念。例如，“虎斑猫”和“埃及猫”这两个概念共享同一个父节点（“家猫”）。分类器很可能更容易混淆两个兄弟类别，因此会给出与真实标签（GT）密切相关的假阳性（FP）预测。因此，与兄弟类别百分比较高的模型相比，兄弟类别数量较少表明分类能力较弱。

解释 在评估阶段提供，旨在回答为什么预训练分类器无法正确分类属于某一类别的未经筛选的图像 c。

假阳性预测包含有关哪些类别会与真实标签混淆的宝贵信息。每类误分类频率（MF）是指每个假阳性类别 f 在假阳性实例总数中所占的百分比。因此，给定一个包含 N 个类别的数据集， c 为真实类别， f 为其中一个假阳性类别，则 $c \to f$ 误分类的误分类频率为：

$$MF_c = \frac{FP_{i=f}}{\sum_{i=0}^{i=N} FP_i} \cdot 100\% \tag{2}$$

MF 可以提取所有 $f \neq c$ 假阳性类别的分数，从而发现影响最大的误分类。较高的 MF 分数表示分类器倾向于将样本判为假阳性类别，而非真实标签，因此表明数据集中可能存在分类器偏差或标注错误。具体而言，如果标注已尽可能准确，那么分类器偏差是指持续将类别 c 的样本分类为类别 f的样本。当然，这一要求并非总能满足，尤其是在需要专家标注员的情况下，例如细粒度分类。另一方面，由于我们的可解释评估方法能够捕捉此类误分类模式，因此无需事先判定误分类的来源。人工标注员可以在后续阶段参与，识别并核实误分类的来源。

4. 实验

在以下所有实验中，我们将每个类别的爬取图像阈值设为 T=50 。我们将展示对随机初始查询的结果，作为概念验证，以说明我们的发现。为此，我们提供查询词“猫”，它会返回以下 WordNet 下位词（也对应 ImageNet 标签）：

H={’安哥拉猫’, ’美洲狮’, ’埃及猫’, ’豹猫’, ’猞猁猫’,’波斯猫’, ’暹罗猫’, ’虎斑猫’, ’虎斑猫’}

只要能够映射到 WordNet，就可以对其他选定的查询词重复进行相同的实验。

4.1. 卷积分类器

我们采用了以下卷积神经网络分类器：VGG16/19, [24], ResNet50/101/152 [26], Incep-tionV3 [25], InceptionResnetV2 [28], Xception [27], MobileNetV2 [49], NasNet-Large [50], DenseNet121/169/201 [51], EfficientNet-B7 [52], ConvNeXt [29]。我们在表 1 中展示了卷积神经网络分类器的结果。加粗实例表示其准确率低于各模型作者分别报告的该模型在 ImageNet 上的最佳准确率[1]。带下划线的单元格表示每个类别的最佳准确率/兄弟类别百分比得分。表 1 中缺少模型或关键词，意味着其对应的准确率分数为零。例如，我们观察到 InceptionV3、InceptionResNetV2、Xception、NASNetLarge、

[1]https://paperswithcode.com/sota/image-classification-on-imagenet

表1使用卷积神经网络的分类结果。粗体条目表示准确率低于最佳模型准确率。

模型	标签	准确率↑	同级类别↑	标签	准确率↑	同类项↑
ResNet50	虎斑猫	**50.00**%	24.00%	暹罗猫	90.00%	0.00%
ResNet101		**52.00**%	41.67%		88.00%	16.67%
ResNet152		**50.00**%	12.00%		90.00%	20.00%
VGG16		**38.00**%	38.71%		82.00%	11.11%
VGG19		**50.00**%	32.00%		88.00%	16.67%
MobileNetV2		**2.00**%	2.04%		**4.00**%	0.00%
EfficientNet		**10.00**%	33.33%		96.00%	100.00%
ConvNeXt		**60.00**%	15.00%		92.00%	75.00%
ResNet50	猞猁猫	82.00%	0.00%	美洲狮	84.00%	0.00%
ResNet101		84.00%	0.00%		78.00%	0.00%
ResNet152		86.00%	0.00%		88.00%	0.00%
VGG16		82.00%	0.00%		86.00%	0.00%
VGG19		80.00%	0.00%		78.00%	0.00%
EfficientNet		90.00%	0.00%		98.00%	100.00%
ConvNeXt		92.00%	0.00%		98.00%	0.00%
ResNet50	虎斑猫	**18.33**%	0.00%	波斯猫	92.00%	25.00%
ResNet101		**23.33**%	0.00%		88.00%	16.67%
ResNet152		**26.67**%	0.00%		88.00%	33.33%
VGG16		**20.00**%	0.00%		86.00%	14.29%
VGG19		**28.33**%	0.00%		80.00%	10.00%
MobileNetV2		**1.67**%	0.00%		**8.00**%	2.17%
EfficientNet		**36.67**%	0.00%		98.00%	100.00%
ConvNeXt		**26.67**%	0.00%		98.00%	100.00%
ResNet50	豹猫	**12.00**%	18.18%	安哥拉猫	**12.00**%	50.00%
ResNet101		**12.00**%	15.91%		**20.00**%	50.00%
ResNet152		**4.00**%	14.58%		**20.00**%	62.50%
VGG16		**10.00**%	6.67%		**10.00**%	46.67%
VGG19		**10.00**%	6.67%		**8.00**%	54.35%
EfficientNet		**2.00**%	16.33%		**4.00**%	95.83%
ConvNeXt		**16.00**%	16.67%		**10.00**%	88.89%
ResNet50	埃及猫	**24.00**%	2.63%	cat	82.05%	0.00%
ResNet101		**30.00**%	2.86%		82.05%	0.00%
ResNet152		**34.00**%	6.06%		79.49%	0.00%
VGG16		**28.00**%	0.00%		87.18%	0.00%
VGG19		**26.00**%	2.70%		76.92%	0.00%
MobileNetV2		**0.00**%	0.00%		**2.56**%	0.00%
EfficientNet		**70.00**%	0.00%		92.31%	0.00%
ConvNeXt		**52.00**%	0.00%		94.87%	0.00%

DenseNet121/169/201，这意味着它们完全无法正确分类爬取图像，即使这些图像属于在使用其他分类器时表现出令人满意准确率的类别。MobileNetV2 在所有类别上的性能也都出现了下降。我们

之后将探讨层次知识能否帮助提取与这种出乎意料的低性能相关的有意义信息。

从表1 还能得出其他结论：某些类别很容易分类（“暹罗猫”“猞猁猫”“美洲狮”“波斯猫”“猫”），而另一些类别则不然（“虎斑猫”“老虎”“埃及猫”“豹猫”“安哥拉猫”）。由于我们对动物物种没有专门知识，将再次借助 WordNet 来解释这种现象。兄弟类别百分比初步反映了在细粒度设置下，相似类别之间的混淆程度。例如，尽管“暹罗猫”和“美洲狮”类别的准确率分数都很高，但两者的兄弟类别百分比表现截然不同：对于“暹罗猫”的真实标签，大多数卷积神经网络分类器都会产生一些兄弟类别假阳性；而“美洲狮”的真实标签则相反，大多不会出现兄弟类别误分类。这表明，对于“暹罗猫”，如果样本被误分类，它很可能属于一个概念相似的类别；而对于“美洲狮”，误分类时的假阳性则属于语义距离较远的类别。

关于模型能力，我们观察到，对于“暹罗猫”和“美洲狮”类别，所有 ResNet50 假阳性都属于非兄弟类别，这与 EfficientNet 的假阳性相反，后者全部属于兄弟类别。通过进一步考察其他类别，我们观察到，总体而言，EfficientNet 相较于 ResNet50 达到了更高的兄弟类别百分比，这意味着与 ResNet50 的误分类相比，EfficientNet 的误分类更为合理。

4.2. 基于Transformer的分类器

使用了以下基于Transformer的图像分类器：ViT [5], Regnet-x [33], DeiT [31],BeiT [9], CLIP [53], Swin Transformer V2 [32]。基于Transformer的分类器的结果见表 2。关于模型难以进行预测的类别，我们发现了类似的模式：属于“虎斑猫”、“虎斑猫”和“埃及猫”类别的样本，其分类准确率低于“暹罗猫”、“猞猁猫”、“美洲狮”、“波斯猫”、“猫”、“安哥拉猫”和“豹猫”。我们推测，这种行为背后可能存在一个共同原因，很可能是由于细粒度分类设置中不可避免的类内相似性所致。

关于模型性能，除了单独评估准确率外，我们还考察兄弟类别百分比。与卷积神经网络模型相比，基于 Transformer 的模型在兄弟类别误分类方面的行为更难解释，因为对于某些类别，模型可能给出较高的兄弟类别百分比，而对于其他类别则给出较低的百分比，反之亦然。例如，与其他模型在相同类别上的表现相比，BeiT 对“虎斑猫”（3.45%）、“暹罗猫”（0%）和“波斯猫”（10%）的兄弟类别百分比得分较低；另一方面，它在“豹猫”（78.72%）、“虎斑猫”（22.45%）和“埃及猫”（22.50%）上的兄弟类别得分最高。关于结果可解释性的更多内容见第 4.3 节。

4.3. 解释误解

在表 3、4 & 5中，我们报告了每个真实类别（GT）和每个模型对应的前3个误分类，以及每个假阳性（FP）标签的误分类频率（MF）。GT列仅指猫科物种，即使省略了“猫”一词（例如，“老虎”GT

表2使用 Transformer 的分类结果。粗体条目表示准确率低于最佳模型准确率，下划线指标表示每个类别的最佳指标表现。

模型	标签	准确率↑	兄弟类↑	标签	准确率↑	同胞↑
ViT	虎斑猫	**44.00%**	42.86%	暹罗猫	92.00%	50.00%
BeiT		**42.00%**	3.45%		94.00%	0.00%
DeiT		**60.00%**	30.00%		94.00%	33.33%
Swin		**48.00%**	30.77%		94.00%	100.00%
xRegNet		**52.00%**	25.00%		92.00%	50.00%
CLIP		**30.00%**	28.57%		96.00%	50.00%
ViT	猞猁猫	90.00%	0.00%	美洲狮	96.00%	0.00%
BeiT		**26.00%**	0.00%		92.00%	0.00%
DeiT		92.00%	0.00%		96.00%	0.00%
Swin		86.00%	0.00%		96.00%	0.00%
xRegNet		90.00%	0.00%		96.00%	50.00%
CLIP		86.00%	0.00%		92.00%	0.00%
ViT	虎斑猫	**18.33%**	0.00%	波斯猫	92.00%	75.00%
BeiT		**18.33%**	22.45%		80.00%	10.00%
DeiT		**15.00%**	0.00%		96.00%	50.00%
Swin		**21.67%**	0.00%		96.00%	50.00%
xRegNet		**35.00%**	0.00%		98.00%	100.00%
CLIP		46.67%	0.00%		96.00%	50.00%
ViT	豹猫	**12.00%**	2.27%	安哥拉猫	**6.00%**	89.36%
BeiT		**6.00%**	78.72%		**62.00%**	52.63%
DeiT		**10.00%**	11.11%		**0.00%**	94.00%
Swin		**14.00%**	9.30%		**8.00%**	95.65%
xRegNet		**6.00%**	21.28%		**0.00%**	76.00%
CLIP		**10.00%**	55.56%		**8.00%**	91.30%
ViT	埃及猫	**38.00%**	3.23%	cat	89.74%	0.00%
BeiT		**20.00%**	22.50%		**53.85%**	0.00%
DeiT		**36.00%**	3.12%		94.87%	0.00%
Swin		**52.00%**	0.00%		89.74%	0.00%
xRegNet		**48.00%**	0.00%		92.31%	0.00%
CLIP		70.00%	6.67%		69.23%	0.00%

条目指的是“虎斑猫”）。我们用红色 突出显示与真实标签在语义上相距较远的无关假阳性类别，而涉及近邻类别的误解则用蓝色突出显示。此外，洋红色 表示某个假阳性实际上是真实类别（GT）的直接（1跳）上位词。受篇幅限制，我们在此展示所有基于 Transformer 的模型，但仅展示总测试卷积神经网络模型中的一部分；更多结果请参见附录。

有趣的是，我们发现了一些令人意外的常见误判，例如将猫科物种与“墨西哥无毛犬”这一犬种混淆。对于卷积神经网络分类器，我们在所有受调查的模型中都发现了这一现象：真实标签为“埃及猫”的样本中，ResNet50 的假阳性有 10.53% 被归入“墨西哥无毛犬”类别；ResNet101 的假阳性中这一比例为 14.29%，ResNet152 的假阳性中则为 18.18%

表3选定真实猫类别的常见误分类及误分类频率（卷积神经网络）。

模型	GT	前1		前2		前3	
		FP	MF	FP	MF	FP	MF
Res Net50	虎斑猫	虎斑猫	32.00%	埃及猫	24.00%	网站	8.00
	安哥拉猫	波斯猫	34.00%	北极狐	11.36%	lynx	9.09%
	lynx	郊狼	22.22%	虎斑猫	11.11%	埃及猫	11.11%
	暹罗猫	大丹犬	20.00%	hare	20.00%	美洲白鹭	20.00%
	老虎	虎斑猫	40.82%	埃及猫	20.41%	老虎	14.29%
	波斯猫	古代英国牧羊犬	25.00%	暹罗猫	25.00%	手斧	25.00%
	美洲狮	lynx	25.00%	马里努阿犬	25.00%	沙袋鼠	25.00%
	豹	埃及猫	30.00%	虎斑猫	16.00%	美洲豹	12.00%
	埃及的cat	墨西哥无毛犬	10.53%	mask	5.26%	漫画书	5.26%
		毛皮大衣	14.29%	纸板箱	14.29%	书籍封套	14.29%
Res Net 101	虎斑猫	埃及猫	41.67%	虎斑猫	29.17%	网站	8.33%
	安哥拉猫	波斯猫	32.50%	埃及猫	12.50	lynx	10.00%
	lynx	虎斑猫	12.50%	埃及猫	12.50%	猎豹	12.50%
	暹罗猫	波士顿斗牛犬	16.67%	埃及猫	16.67%	hare	16.67%
	老虎	虎斑猫	34.78%	虎斑猫	17.39%	埃及猫	15.22%
	波斯猫	荷兰毛狮犬	16.67%	豚鼠	16.67%	柯利牧羊犬	16.67%
	美洲狮	lynx	45.45%	狐獴	9.09%	豺	9.09%
	豹	埃及猫	36.00%	虎斑猫	14.00%	豹	12.00%
	埃及的cat	墨西哥无毛犬	14.29%	mask	8.57%	海狮	5.71%
		猕猴	14.29%	理发店	14.29%	彭布罗克威尔士柯基犬	14.29%
Res Net 152	虎斑猫	虎斑猫	40.00%	埃及猫	12.00%	lynx	12.00%
	安哥拉猫	波斯猫	35.00%	暹罗猫	10.00%	淋浴帷幕	10.00%
	lynx	虎斑猫	42.86%	郊狼	14.29%	诺里奇梗	14.29
	暹罗猫	惠比特犬	20.00%	埃及猫	20.00%	安哥拉猫	20.00%
	老虎	虎斑猫	34.09%	埃及猫	18.18%	老虎	15.91%
	波斯猫	暹罗猫	33.33%	柯利牧羊犬	16.67%	毛皮大衣	16.67%
	美洲狮	menu	16.67%	野猪	16.67%	沙袋鼠	16.67%
	豹	埃及猫	28.00%	lynx	22.00%	美洲豹	16.00%
	埃及的cat	墨西哥无毛犬	18.18%	网站	9.09%	虎斑猫	6.06%
		猕猴	12.50%	彭布罗克威尔士柯基犬	12.50%	吉娃娃	12.50%
VGG 16	虎斑猫	埃及猫	38.71%	虎斑猫	22.58%	林兔	3.23%
	安哥拉猫	波斯猫	26.67%	埃及猫	15.56%	lynx	8.89%
	lynx	郊狼	33.33	埃及猫	22.22%	马达加斯加毛皮大衣	11.11%
	暹罗猫	墨西哥无毛犬	22.22%	惠比特犬	11.11%		11.11%
	老虎	虎斑猫	33.33%	埃及猫	20.83%	老虎	16.67%
	波斯猫	北极狐	14.29%	安哥拉猫	14.29%	lynx	14.29%
	美洲狮	lynx	42.86%	郊狼	28.57%	menu	14.29%
	豹	埃及猫	42.00%	lynx	18.00%	美洲豹	10.00%
	埃及的cat	墨西哥无毛犬	8.33%	lynx	5.56%	宽边帽	5.56%
		诺里奇梗	20.00%	舒柏奇犬	20.00%	敏狐	20.00%

表4选定真实猫类别的常见误分类及其误分类频率（Transformer）。

模型	GT	前1 FP	前1 MF	前2 FP	前2 MF	前3 FP	前3 MF
CLIP	虎斑猫	马达加斯加	40.00%	埃及猫	22.86%	虎斑猫	11.43
	安哥拉猫	波斯猫	78.26%	马达加斯加	6.52%	暹罗猫	6.52%
	lynx	马达加斯加	14.29%	豹猫	14.29%	灰狐	14.29%
	暹罗猫	臭鼬	50.00%	波斯猫	50.00%	-	-
	老虎	埃及猫	30.77%	马达加斯加	19.23%	豹猫	15.38%
	波斯猫	马达加斯加	50.00%	暹罗猫	50.00%	-	-
	美洲狮	lynx	75.00%	马达加斯加	25.00%	-	-
	豹	虎斑猫	55.56%	马达加斯加	17.78%	埃及猫	11.11%
	埃及的cat	墨西哥无毛犬	26.67%	马达加斯加	26.67%	狐狸	6.67%
		马达加斯加	66.67%	橙色	16.67%	bib	8.33%
BeiT	虎斑猫	虎斑猫	17.24%	cat	13.79%	家养	13.79%
	安哥拉猫	波斯猫	42.11%	家养	21.05%	四足动物	5.26%
	lynx	欧亚猞猁	59.46%	加拿大猞猁	16.22%	短尾猫	5.41%
	暹罗猫	小猫	66.67%	猫科动物	33.33%	-	-
	老虎	虎斑猫	23.40%	长尾虎猫	14.89%	家养	6.38%
	波斯猫	家养	20.00%	安哥拉猫	10.00%	养家糊口的人	10.00%
	美洲狮	猫科动物	25.00%	大型猫科动物	25.00%	cub	25.00%
	豹	长尾虎猫	42.55%	虎猫	21.28%	斑点猞猁	8.51%
	埃及的cat	阿比西尼亚猫	15.00%	墨西哥无毛犬	10.00%	捕鼠者	5.00%
		猫科动物	33.33%	小猫	22.22%	餐饮承办人	11.11%
DeiT	虎斑猫	虎斑猫	35.00%	埃及猫	30.00%	网站	15.00%
	安哥拉猫	波斯猫	62.00%	埃及猫	28.00%	虎斑猫	2.00%
	lynx	虎斑猫	75.00%	郊狼	25.00%	-	-
	暹罗猫	埃及猫	33.33%	墨西哥无毛犬	33.33%	lynx	33.33%
	老虎	虎斑猫	37.50%	埃及猫	27.50%	豹猫	12.50%
	波斯猫	软毛麦色梗	50.00%	暹罗猫	50.00%	-	-
	美洲狮	网站	50.00%	澳洲野犬	50.00%	-	-
	豹	埃及猫	48.89%	lynx	22.22%	虎斑猫	11.11%
	埃及的cat	墨西哥无毛犬	15.62%	漫画书	9.38%	凯尔派犬	3.12%
		毛皮大衣	50.00%	吉娃娃	50.00%	-	-
xRegNet	虎斑猫	虎斑猫	62.50%	埃及猫	20.83%	menu	4.17%
	安哥拉猫	波斯猫	48.00%	埃及猫	18.00%	lynx	8.00%
	lynx	虎斑猫	40.00%	虎斑猫	20.00%	埃及猫	20.00%
	暹罗猫	埃及猫	50.00%	臭鼬	25.00%	lynx	25.00%
	老虎	虎斑猫	40.00%	埃及猫	25.71%	lynx	11.43%
	波斯猫	暹罗猫	100.0%	-	-	-	-
	美洲狮	虎斑猫	50.00%	lynx	50.00%	-	-
	豹	埃及猫	40.43%	虎斑猫	21.28%	lynx	17.02%
	埃及的cat	墨西哥无毛犬	15.38%	mask	7.69%	漫画书	7.69%
		漫画书	33.33%	tub	33.33%	雄鸭	33.33%

表5（续表4）。常见误分类及误分类频率。

模型	GT	前1 FP	前1 MF	前2 FP	前2 MF	前3 FP	前3 MF
Swin	虎斑猫	虎斑猫	57.69%	埃及猫	30.77%	网站	7.69%
	安哥拉猫	波斯猫	58.70%	埃及猫	26.09%	虎斑猫	10.87%
	lynx	虎斑猫	57.14%	毛皮大衣	14.29%	森林狼	14.29%
	暹罗猫	埃及猫	100.0%	-	-	-	-
	老虎	虎斑猫	35.00%	埃及猫	32.50%	豹猫	12.50%
	波斯猫	暹罗猫	50.00%	手持吹风机	50.00%	-	-
	美洲狮	网站	50.00%	爱尔兰语	50.00%	-	-
	豹	埃及猫	44.19%	猎狼犬 lynx	37.21%	虎斑猫	9.30%
	埃及的 cat	墨西哥无毛犬 毛皮大衣	20.83% 25.00%	漫画书 运动衫	16.67% 25.00%	台灯 吉娃娃	8.33% 25.00%
ViT	虎斑猫	埃及猫	42.86%	虎斑猫	32.14%	网站	10.71%
	安哥拉猫	埃及猫	48.94%	波斯猫	38.30%	虎斑猫	2.13%
	lynx	虎斑猫	40.00%	埃及猫	40.00%	森林狼	20.00%
	暹罗猫	埃及猫	50.00%	吉娃娃	25.00%	墨西哥无毛犬	25.00%
	老虎	埃及猫	48.78%	虎斑猫	26.83%	豹猫	14.63%
	波斯猫	塑料袋	25.00%	埃及猫	25.00%	暹罗猫	25.00%
	美洲狮	埃及猫	50.00%	马里努阿犬	50.00%	-	-
	豹	埃及猫	86.36%	雪豹	2.27%	网站	2.27%
	埃及的 cat	墨西哥无毛犬	16.13%	基座	12.90%	vase	6.45%
		洗衣机	25.00%	毛皮大衣	25.00%	墨西哥无毛犬	25.00%

以及VGG16误报中的8.33%。此外，’wallaby’、’jaguar’、’sea lion’、’cheetah’、’arctic fox’和’coyote’等更多动物也经常被预测为假阳性。

对于Transformer模型，当提供’egyptian cat’的真实标签时，所有分类器都观察到了’egyptian cat’ →’mexican hairless’异常现象，从而得到以下’ mex ican hairless’假阳性百分比：CLIP为26.67%，BeiT为10%，DeiT为15.62%，xRegNet为15.38%，Swin为20.83%，ViT为16.33%。显然，无论使用卷积神经网络分类器还是Transformer分类器，’egyptian cat’的图像都经常被错误地识别为’mexican hairless dogs’。对’egyptian cat’图像和’mexican hairless dog’图像进行定性分析表明，尽管这些动物具有相似的耳朵形状以及相当无毛、瘦削的身体，但它们显然并不相同。因此，我们可以认为，基于Transformer的分类器偏向于纹理，这验证了针对卷积神经网络报告的相关观察结果 [11]。此外，耳朵形状充当了混杂因素，掩盖了其他实际上具有明显差异的动物特征。还存在涉及’armadillo’、’chihuahua’、’soft-coated wheaten terrier’、’kelpie’等动物的更多误分类。

更令人惊讶的是，有些误分类甚至不涉及动物物种。例如，卷积神经网络分类器将“虎斑猫”预测为“网站”，将“波斯猫”预测为“孵化的”，将“猫”预测为“理发店”，将“美洲狮”预测为“菜单”等。所有 ResNet50/101/152 和 VGG16 至少都会出现一次此类误分类，这不禁让人高度质疑猫科物种的哪些特征

促成了这样的预测。

使用 Transformer 时涉及非动物类别的误分类（表格4、5）呈现出以下有趣的异常现象：使用 DeiT 时，50% 的假阳性样本中，‘猫’被分类为‘毛皮大衣’。这一不可忽视的误分类率再次证实了前文所述的纹理偏差。类似地，xRegNet 会将‘埃及猫’图像分别误分类为‘面具’和‘漫画书’，在假阳性样本中的占比均为 7.69%。这些类别也曾出现在卷积神经网络的误分类中。我们无法对‘面具’这一误分类给出人类可理解的解释，因为‘面具’一词可能指代许多不同的物体。我们推测，ImageNet 中标注为‘面具’的图像可能包含外观与猫相似的狂欢节面具，因此缺少上下文使 xRegNet 产生了混淆。DeiT 将‘埃及猫’图像误分类为‘漫画书’的情况占 9.38%；xRegNet 将‘猫’照片误分类为‘漫画书’的情况占 33.33%；Swin 将‘埃及猫’误分类为‘漫画书’的情况占 16.67%。这可能是因为爬取的图像中包含卡通风格图像，难以明确地将其视为猫。其他涉及无关类别的有趣误分类还包括：‘猫’→‘洗衣机’（使用 ViT 时占假阳性的 25%）、‘豹猫’→‘网站’（使用 ViT 时占假阳性的 2.27%，使用 DeiT 时占 15%）、‘波斯猫’→‘塑料袋’（使用 ViT 时占假阳性的 25%）、‘猫’→‘运动衫’（使用 Swin 时占假阳性的 25%）、‘埃及猫’→‘台灯’（使用 Swin 时占假阳性的 8.33%）、‘猫’→‘浴缸’（使用 xRegNet 时占假阳性的 33.33%），等等。

一个有趣的观察围绕“埃及猫”标签展开。对于卷积神经网络模型，“埃及猫”真实标签排名前3的假阳性几乎都对应于无关的 ImageNet 类别。相比之下，“虎斑猫”、“安哥拉猫”和“虎斑猫”呈现出更合理的假阳性，通常涉及同类类别（以蓝色突出显示）。对于 Transformer 模型，我们观察到，“埃及猫”标签总是至少与一个无关的 ImageNet 类别混淆，而“安哥拉猫”只会与其他猫科物种混淆，不会与概念上相距甚远的类别混淆。因此，“埃及猫”的爬取图像似乎包含一些会频繁干扰分类过程的误导性视觉特征。确实，查看“埃及猫”的爬取图像时，其中一些是猫主题纪念品的绘画或照片；然而，“台灯”或“犰狳”之类的误解无法由人工检查者从视觉上解释，这引出了更多相关问题。卷积神经网络分类器（表 3）与基于Transformer的分类器（表 4、5）的比较表明，Transformer更擅长检索与真实标签相似的类别；观察卷积神经网络中以红色突出显示的无关误分类数量高于Transformer的结果，这一点显而易见。

结合表 3、4 & 5 与表 1& 2，我们可以得出一些非常有趣的发现：较低的分类指标分数与误分类类别之间的相关性如何？我们首先考察准确率分数较低的类别（“虎斑猫”“虎斑猫”“埃及猫”），并将它们与经常出现无关误分类的类别（“埃及猫”和“猫”，其次是“虎斑猫”和“猞猁”）进行比较。对“埃及猫”图像进行分类时，既会得到较低的分类分数，也会产生无关的假阳性。另一方面，尽管“猫”图像的准确率分数较高，但一旦发生误分类，结果往往与其高度无关。“虎斑猫”的准确率较低，不过误分类尚算合理，因为模型会预测为其他猫科物种。令人意外的是，“虎斑猫”的兄弟类别百分比也较低，这表明假阳性并非直接与真实标签中的“虎斑猫”类别相关。在这种情况下，我们认为假阳性（“埃及猫”“虎斑猫”“豹猫”等）属于“虎斑猫”概念类别中关系较远的类别，尽管它们具有一些相似特征。

总体而言，在整个分析过程中，我们证明了分类准确率无法揭示

分类器行为背后的全部真相；为此，知识来源能够帮助我们了解这一过程的内在运作机制。通过分析一组相关的 ImageNet 标签（猫科物种），我们已经将分类准确率与分类相关性区分开来：假阳性可能与真实标签高度相关（例如将其误分类为“虎斑猫”），也可能无关（例如误分类为“猫”）。因此，我们认为，细粒度分类还需要 细粒度评估；在知识的驱动下，这种评估能够提供富有洞察力的信息。表4、5 所呈现的

表6基于使用卷积神经网络分类器的WordNet距离的概念度量指标。

模型	标签	路径相似度↑	LCH↑	WUPS↑	标签	路径 sim↑	LCH↑	WUPS↑
ResNet50		0.18	1.79	0.69		0.10	1.25	0.57
ResNet101		0.22	1.99	0.75		0.15	1.60	0.72
ResNet152	虎斑猫	0.16	1.59	0.62	暹罗猫	0.17	1.71	0.67
VGG16	cat	0.21	1.85	0.70	cat	0.16	1.65	0.65
VGG19		0.18	1.68	0.63		0.17	1.73	0.70
MobileNetV2		0.09	1.13	0.39		0.09	1.15	0.40
EfficientNet		**0.24**	**2.17**	**0.86**		**0.33**	**2.54**	**0.88**
ResNet50		0.05	0.53	0.11		0.08	0.97	0.49
ResNet101		0.05	**0.62**	**0.13**		0.08	0.90	0.41
ResNet152	猞猁猫	0.05	0.56	0.09	美洲狮	0.08	1.01	0.49
VGG16		0.04	0.46	0.08	cat	0.08	0.88	0.37
VGG19		0.04	0.48	0.08		0.07	0.86	0.38
EfficientNet		0.05	0.54	0.09		**0.33**	**2.54**	**0.94**
ResNet50		**0.15**	**1.61**	**0.70**		0.16	1.59	0.60
ResNet101		0.14	1.47	0.64		0.17	1.73	0.66
ResNet152	虎斑猫	0.13	1.43	0.61	波斯猫	0.17	1.57	0.56
VGG16		0.14	1.51	0.65	cat	0.12	1.27	0.50
VGG19		0.13	1.43	0.61		0.13	1.45	0.57
MobileNetV2		0.07	0.90	0.41		0.09	1.19	0.43
EfficientNet		0.13	1.41	0.60		**0.33**	**2.54**	**0.88**
ResNet50		0.17	1.62	0.67		0.22	1.94	0.72
ResNet101		0.17	1.62	0.67		0.22	1.93	0.71
ResNet152	豹	0.16	1.55	0.64	安哥拉猫	0.24	2.05	0.73
VGG16	cat	0.15	1.49	0.62	cat	0.21	1.90	0.72
VGG19		0.15	1.53	0.64		0.23	2.01	0.75
EfficientNet		**0.18**	**1.68**	**0.68**		**0.32**	**2.48**	**0.86**
ResNet50		0.11	1.32	0.51		0.11	1.29	0.56
ResNet101		0.11	1.32	0.50		0.11	1.34	0.63
ResNet152	埃及的	**0.12**	**1.40**	**0.55**	cat	0.11	1.35	0.61
VGG16	cat	0.09	1.15	0.41		**0.14**	**1.67**	**0.80**
VGG19		0.10	1.21	0.45		0.12	1.42	0.64
MobileNetV2		0.08	1.06	0.36		0.07	0.88	0.34
EfficientNet		0.10	1.24	0.49		0.12	1.52	0.74

将在下一章节中进行量化和验证。

4.4 知识驱动指标

前述关于需要进行细粒度评估的论点，可通过展示基于知识驱动指标的结果来支持，这些指标基于 WordNet 提供的概念距离（表6和7）。由于较高的路径相似度/LCH、WUPS 分数更好，我们用粗体标出每个类别中最佳（最高）的分数。

通过比较各类别的路径相似度、LCH 和 WUPS 指标，我们发现，包含大量无关假阳性（在表格中以红色标记）的类别（例如「美洲狮4、5」）

表 7（表6的续表）。使用 Transformer 基于 WordNet距离的概念度量指标。

模型	标签	路径相似度↑	LCH↑	WUPS↑	标签	路径相似度↑	LCH↑	WUPS↑
ViT		0.23	2.02	0.76		0.23	2.01	0.71
BeiT		**0.24**	2.02	0.75		0.18	1.88	0.77
DeiT	虎斑猫	0.20	1.83	0.70	暹罗猫	0.19	1.72	0.58
Swin	cat	0.23	**2.07**	**0.82**	cat	**0.33**	**2.54**	**0.88**
xRegNet		0.22	2.00	0.79		0.21	1.84	0.66
CLIP		0.18	1.78	0.75		0.24	2.12	0.84
ViT		0.05	0.55	0.09		0.15	1.63	**0.79**
BeiT		0.04	0.33	0.07		**0.21**	**1.94**	**0.79**
DeiT	猞猁猫	0.05	0.54	0.09	美洲狮	0.09	1.13	0.54
Swin		0.05	**0.56**	0.09	cat	0.07	0.97	0.51
xRegNet		0.04	0.50	0.08		0.19	1.44	0.50
CLIP		0.04	0.51	**0.10**		0.06	0.62	0.24
ViT		0.15	1.60	0.69		0.27	2.19	0.76
BeiT		**0.21**	**1.89**	**0.76**		0.22	1.87	0.66
DeiT	虎斑猫	0.13	1.42	0.61	波斯猫	0.24	2.12	0.80
Swin		0.14	1.47	0.63	cat	0.21	1.85	0.65
xRegNet		0.14	1.50	0.64		**0.33**	**2.54**	**0.88**
CLIP		0.12	1.39	0.62		0.22	1.99	0.80
ViT		0.17	1.76	0.75		0.31	2.39	0.83
BeiT		**0.31**	**2.47**	**0.93**		0.30	2.22	0.73
DeiT	豹	0.15	1.47	0.60	安哥拉猫	**0.32**	2.48	**0.86**
Swin	cat	0.13	1.26	0.50	cat	**0.32**	**2.49**	**0.86**
xRegNet		0.18	1.70	0.69		0.28	2.22	0.78
CLIP		0.22	1.87	0.74		0.31	2.44	**0.86**
ViT		0.11	1.33	0.50		0.09	1.15	0.50
BeiT		**0.16**	1.63	0.60		**0.23**	**1.81**	**0.68**
DeiT	埃及的	0.11	1.31	0.50	cat	0.05	0.72	0.29
Swin	cat	0.11	1.34	0.52		0.05	0.73	0.29
xRegNet		0.10	1.24	0.46		0.06	0.89	0.39
CLIP		0.15	**1.65**	**0.70**		0.12	1.42	0.66

猫」和「猞猁猫」，其次是「埃及猫」和「猫」，在表格6、7」中也呈现出较低的知识驱动指标得分，这符合预期。其他类别，例如「安哥拉猫」「豹猫」和「虎斑猫」，其误分类来自相关类别（兄弟类别或父类别），也呈现出较高的知识驱动指标得分。因此，我们可以放心地认为，使用知识驱动指标评估细粒度分类结果，与人类可理解的相似性概念高度相关，因此值得信赖。

考察卷积神经网络分类器时，模型性能的差异相当明显。对于大多数类别，与其他分类器相比，EfficientNet 更擅长预测相关性更高的假阳性图像。另一方面，对于基于 Transformer 的分类器，则较难得出类似结论，因为不同模型在不同类别上的表现更好；不过，与卷积神经网络分类器相比，大多数类别的知识驱动指标结果相同或更高。尽管这种差异并不显著，但基于 Transformer 的模型展现出了更强的能力：在未能预测出真实标签时，能够预测出相关性更高的类别。

5. 结论

在本研究中，我们构建了一种涉及未经筛选的网络图像的新型分布偏移，并基于此测试了卷积和基于 Transformer 的图像分类器。分类时选择密切相关的类别，由层次知识提供指导；我们再次运用这种知识来评估结果质量。我们证明，准确率相关指标只能揭示分类评估的表层情况，因为它们无法捕捉误分类样本与真实标签之间的语义关系。为此，我们提出了一种可解释的知识驱动评估方案，通过提供假阳性与真实标签之间的语义距离，量化误分类相关性。我们还利用同一方案，在所构建的分布偏移下比较卷积神经网络与基于 Transformer 的模型的分类能力。作为未来工作，我们计划将分析扩展到更多查询词，以考察当前发现的适用范围；同时，将未经筛选图像的分类设置与人工损坏相结合，以深化我们的认识。

致谢

本研究工作得到希腊研究与创新基金会（HFRI）HFRI第三次博士奖学金项目（奖学金编号5537）的资助。

参考文献

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei，Imagenet：大规模层次化图像数据库，载于：2009年 IEEE计算机视觉与模式识别会议，2009年，第248–255页。doi:10.1109/CVPR.2009.5206848.[2] A. Krizhevsky, I. Sutskever, G. E. Hinton，使用深度卷积神经网络进行Imagenet分类，载于：《第25届神经信息处理系统进展国际会议论文集》第1卷，NIPS'12，Curran Associates Inc.，美国纽约州雷德胡克，2012年，第1097–1105页。

[3] Z. Li, W. Yang, S. Peng, F. Liu，卷积神经网络综述：分析、应用与展望，arXiv，2020年。URL：https://arxiv.org/abs/2004.02806。doi：10.48550/ARXIV.2004.02806.[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin，注意力机制就是一切，2017年。URL：https://arxiv.org/abs/1706.03762。doi:10.48550/ARXIV.1706.03762.[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. De- hghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby，一张图像胜过16×16个词：用于大规模图像识别的Transformer，2020年。URL：https://arxiv.org/abs/2010.11929。doi:10.48550/ARXIV.2010.11929.[6] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, Y. Wu，Coca：对比式图像描述器是图像文本基础模型，2022年。URL：https://arxiv.org/abs/2205.01917。doi:10.48550/ARXIV.2205.01917.[7] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, L. Schmidt，模型汤：对多个微调模型的权重进行平均，可在不增加推理时间的情况下提高准确率，载于：K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato（编），第39届国际机器学习会议论文集，机器学习研究会议录第162卷，PMLR，2022年，第23965–23998页。URL：https://proceedings.mlr.press/v162/wortsman22a.html。[8] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo，Swin Transformer V2：扩大容量与分辨率（2021年）。URL：https://arxiv.org/abs/2111.09883。doi:10.48550/ARXIV.2111.09883.[9] H. Bao, L. Dong, S. Piao, F. Wei，BeiT：图像Transformer的BERT预训练，2021年。URL：https://arxiv.org/abs/2106.08254。doi:10.48550/ARXIV.2106.08254.[10] J. Cohen, E. Rosenfeld, Z. Kolter，通过随机平滑实现可认证对抗鲁棒性，载于：K. Chaudhuri, R. Salakhutdinov（编），第36届国际机器学习会议论文集，机器学习研究会议录第97卷，PMLR，2019年，第1310–1320页。URL：https://proceedings.mlr.press/v97/cohen19c.html。[11] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, W. Brendel，基于ImageNet训练的卷积神经网络偏向纹理；增强形状偏置可提高准确率和鲁棒性，2018年。URL：https://arxiv.org/abs/1811.12231。doi:10.48550/ARXIV.1811.12231。[12] F. Yang, Z. Wang, C. Heinze-Deml，使用最坏情况变换进行诱导不变性的正则化，足以提升准确率和空间鲁棒性，载于：H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett（编），《神经信息处理系统进展》第32卷，Curran Associates, Inc.，2019年。URL：https://proceedings.neurips.cc/paper/2019/file/1d01bd2e16f57892f0954902899f0692-Paper.pdf。[13] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, L. Beyer, O. Bachem, M. Tschannen, M. Michalski, O. Bousquet, S. Gelly, N. Houlsby，使用视觉任务适应基准开展的表示学习大规模研究，2019年。URL：https://arxiv.org/abs/1910.04867。doi：10.48550/ARXIV.1910.04867.[14] D. Hendrycks, T. Dietterich，对神经网络应对常见损坏的鲁棒性进行基准测试

以及扰动，2019年。 arXiv:1903.12261.[15] R. Taori、A. Dave、V. Shankar、N. Carlini、B. Recht、L. Schmidt，《衡量图像分类对自然分布偏移的鲁棒性》，2020年。URL：https://arxiv.org/abs/2007. 00644。doi:10.48550/ARXIV.2007.00644.[16] D. Hendrycks、S. Basart、N. Mu、S. Kadavath、F. Wang、E. Dorundo、R. Desai、T. L. Zhu、S. Parajuli、M. Guo、D. X. Song、J. Steinhardt、J. Gilmer，《鲁棒性的多种面貌：对分布外泛化的批判性分析》，2021年IEEE/CVF国际计算机视觉会议（ICCV）（2020）8320–8329。[17]A. Barbu、D. Mayo、J. Alverio、W. Luo、C. Wang、D. Gutfreund、J. Tenenbaum、B. Katz，《ObjectNet：用于推动目标识别模型极限的大规模偏差受控数据集》，载于：H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché-Buc、E. Fox、R. Garnett（编），《神经信息处理系统进展》，第32卷，Curran Associates, Inc.，2019年。URL：https://proceedings.neurips.cc/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf。[18] D. Hendrycks、K. Zhao、S. Basart、J. Steinhardt、D. Song，《自然对抗样本》，2019年。URL：https://arxiv.org/abs/1907.07174。doi:10.48550/ARXIV.1907.07174.[19] T. DeVries、G. W. Taylor，《使用Cutout改进卷积神经网络的正则化》，2017年。URL：https://arxiv.org/abs/1708.04552。doi:10.48550/ARXIV.1708.04552.[20] S. Zheng、Y. Song、T. Leung、I. Goodfellow，《通过稳定性训练提高深度神经网络的鲁棒性》，2016年。URL：https://arxiv.org/abs/1604.04326。doi:10.48550/ARXIV.1604.04326.[21] L. Taylor、G. Nitschke，《使用通用数据增强改进深度学习》，2017年。URL：https://arxiv.org/abs/1708.06020。doi:10.48550/ARXIV.1708.06020.[22] S.-A. Rebuffi、S. Gowal、D. A. Calian、F. Stimberg、O. Wiles、T. Mann，《数据增强可以提高鲁棒性》，2021年。URL：https://arxiv.org/abs/2111.05328。doi:10.48550/ARXIV.2111.05328。[23] C. Fellbaum，《WordNet：电子词汇数据库》（1998年）。[24] K. Simonyan、A. Zisserman，《用于大规模图像识别的超深卷积网络》，CoRR abs/1409.1556（2014年）。[25] C. Szegedy、V. Vanhoucke、S. Ioffe、J. Shlens、Z. Wojna，《重新思考用于计算机视觉的Inception架构》，载于：2016年IEEE计算机视觉与模式识别会议（CVPR），2016年，第2818–2826页。doi:10.1109/CVPR.2016.308.[26] K. He、X. Zhang、S. Ren、J. Sun，《用于图像识别的深度残差学习》，载于：2016年IEEE计算机视觉与模式识别会议（CVPR），2016年，第770–778页。doi:10.1109/CVPR.2016.90.[27] F. Chollet，《Xception：使用深度可分离卷积的深度学习》，2017年IEEE计算机视觉与模式识别会议（CVPR）（2016年）1800–1807。[28] C. Szegedy、S. Ioffe、V. Vanhoucke、A. A. Alemi，《Inception-v4、Inception-ResNet以及残差连接对学习的影响》，载于：第三十一届AAAI人工智能会议论文集，AAAI'17，AAAI Press，2017年，第4278–4284页。[29] Z. Liu、H. Mao、C. Wu、C. Feichtenhofer、T. Darrell、S. Xie，《面向2020年代的卷积网络》，CoRR abs/2201.03545（2022年）。URL：https://arxiv.org/abs/2201.03545。 arXiv:2201.03545.[30] S. Khan、M. Naseer、M. Hayat、S. W. Zamir、F. S. Khan、M. Shah，《视觉中的Transformer：综述》，ACM计算调查54（2022年）1–41。URL：https://doi.org/10.1145%2F3505244。

doi:10.1145/3505244.[31] H. Touvron、M. Cord、M. Douze、F. Massa、A. Sablayrolles、H. Jegou，《通过注意力进行高效图像Transformer训练数据蒸馏》，载于：M. Meila、T. Zhang（编），《第38届国际机器学习会议论文集》，第139卷，机器学习研究会议录，PMLR，2021年，第10347–10357页。[32] Z. Liu、H. Hu、Y. Lin、Z. Yao、Z. Xie、Y. Wei、J. Ning、Y. Cao、Z. Zhang、L. Dong、F. Wei、B. Guo，《Swin Transformer V2：扩展容量和分辨率》，载于：2022年IEEE/CVF计算机视觉与模式识别会议（CVPR），2022年，第11999–12009页。doi:10.1109/CVPR52688.2022.01170.[33] I. Radosavovic、R. P. Kosaraju、R. Girshick、K. He、P. Dollár，《设计网络设计空间》，2020年。[34] S. Paul、P.-Y. Chen，《视觉Transformer是鲁棒学习器》，载于：AAAI人工智能会议，2021年。[35] B. Recht、R. Roelofs、L. Schmidt、V. Shankar，《ImageNet分类器能泛化到ImageNet吗？》，载于：国际机器学习会议，2019年。[36] R. Geirhos、C. R. M. Temme、J. Rauber、H. H. Schütt、M. Bethge、F. A. Wichmann，《人类和深度神经网络中的泛化》，2018年。URL：https://arxiv.org/abs/1808.08750。doi:10.48550/ARXIV.1808.08750.[37] A. Laugros、A. Caplier、M. Ospici，《使用合成扰动衡量对自然分布偏移的鲁棒性》，ArXiv abs/2107.12052（2021年）。[38] D. Hendrycks、K. Zhao、S. Basart、J. Steinhardt、D. Song，《自然对抗样本》，2019年。URL：https://arxiv.org/abs/1907.07174。doi:10.48550/ARXIV.1907.07174.[39] B. Biggio、I. Corona、D. Maiorca、B. Nelson、N. Š rndić、P. Laskov、G. Giacinto、F. Roli，《测试时针对机器学习的规避攻击》，载于：高级信息系统工程，Springer Berlin Heidelberg，2013年，第387–402页。URL：https://doi.org/10.1007%2F978-3-642-40994-3_25。doi:10.1007/978-3-642-40994-3_25.[40] Y. Dong、Q.-A. Fu、X. Yang、T. Pang、H. Su、Z. Xiao、J. Zhu，《图像分类对抗鲁棒性基准测试》，载于：2020年IEEE/CVF计算机视觉与模式识别会议（CVPR），2020年，第318–328页。doi:10.1109/CVPR42600.2020.00040.[41] D. Tsipras、S. Santurkar、L. Engstrom、A. Turner、A. Madry，《鲁棒性可能与准确率相悖》，2018年。URL：https://arxiv.org/abs/1805.12152。doi:10.48550/ARXIV.1805.12152。[42] U. Ozbulak、E. T. Anzaku、W. D. Neve、A. V. Messem，《源图像的选择会严重影响对抗攻击的有效性》，ArXiv abs/2106.07141（2021年）。[43] Y. Wang、E. Ullah、P. Mianjy、R. Arora，《对抗鲁棒性与惰性训练相悖》，ArXiv abs/2207.00411（2022年）。[44] F. Pinto、P. H. S. Torr、P. K. Dokania，《对卷积神经网络与Transformer鲁棒性竞赛的公正看法》，载于：欧洲计算机视觉会议，2022年。[45] Z. Wang、Y. Bai、Y. Zhou、C. Xie，《卷积神经网络能否比Transformer更鲁棒？》，ArXiv abs/2206.03452（2022年）。[46] S. Bhojanapalli、A. Chakrabarti、D. Glasner、D. Li、T. Unterthiner、A. Veit，《理解Transformer用于图像分类的鲁棒性》，2021年。[47] W. Deng、S. Gould、L. Zheng，《模型不变性与泛化之间的强相关性》，ArXiv abs/2207.07065（2022年）。

[48] E. Mintun, A. Kirillov, S. Xie, 《自然腐蚀鲁棒性中增强与腐蚀之间的相互作用》，载于：M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan（编），神经信息处理系统进展，第34卷，Curran Associates, Inc.，2021年，第3571–3583页。网址：https://proceedings.neurips.cc/paper/2021/file/1d49780520898fe37f0cd6b41c5311bf-Paper.pdf。[49] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, 《MobileNetV2：倒置残差与线性瓶颈》，2018年，第4510–4520页。doi：10.1109/CVPR.2018.00474。[50] B. Zoph, V. Vasudevan, J. Shlens, Q. Le, 《学习可迁移架构以实现可扩展图像识别》，2018年，第8697–8710页。doi：10.1109/CVPR.2018.00907。[51] G. Huang, Z. Liu, L. van der Maaten, K. Weinberger, 《密集连接卷积网络》，2017年。doi：10.1109/CVPR.2017.243。[52] M. Tan, Q. Le, 《EfficientNet：重新思考卷积神经网络的模型缩放》，2019年。[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, 《从自然语言监督中学习可迁移视觉模型》，载于：国际机器学习会议，2021年。

A. 更多卷积神经网络误分类

表 8 展示了表 3 中结果的后续部分，涉及其余准确率非零的卷积神经网络模型。显然，分类器的容量在识别相关假阳性方面起着重要作用：MobileNetV2 的准确率分数已经很低，也无法检索出语义相关的假阳性类别。从与该模型对应的大量红色条目中，可以很容易地看出这一点。

除此之外，结果与表 3 中分析的观察结果一致：“埃及猫”标签出现了许多无关的假阳性，而“虎斑猫”或“虎斑猫”标签则并非如此。

表8（表3的续篇）。选定真实猫类别的常见误分类及卷积神经网络的误分类频率。

模型	GT	前1 FP	前1 MF	前2 FP	前2 MF	前3 FP	前3 MF
VGG 19	虎斑猫	埃及猫	28.00%	虎斑猫	20.00%	lynx	16.00%
	安哥拉猫	波斯猫	34.78%	北极狐	10.87%	埃及猫	10.87%
	lynx	埃及猫	20.00%	郊狼	20.00%	森林狼	20.00%
	暹罗猫	惠比特犬	16.67%	毛皮大衣	16.67%	埃及猫	16.67%
	老虎	虎斑猫	34.88%	埃及猫	16.28%	老虎	13.95%
	波斯猫	lynx	20.00%	北京犬	25.00%	毛皮大衣	10.00%
	美洲狮	lynx	45.45%	郊狼	18.18%	森林狼	9.09%
	豹	埃及猫	46.00%	lynx	16.00%	美洲豹	12.00%
	埃及的cat	lynx	8.11%	mask	5.41%	书籍封套	5.41%
		毛皮大衣	11.11%	雪豹	11.11%	捕鼠器	11.11%
移动端 Net V2	虎斑猫	漫画书	14.29%	mask	10.20%	sock	8.16%
	安哥拉猫	淋浴帘	20.00%	纱窗	16.00	聚光灯	8.00%
	lynx	西高地	8.00%	老虎	6.00%	交通信号灯	6.00%
	暹罗猫	白梗 淋浴帘	14.58%	sock	14.58%	mask	14.58%
	老虎	斑马	11.86%	mask	6.78%	maze	6.78%
	波斯猫	聚光灯	10.87%	淋浴帘	8.70%	ant	8.70%
	美洲狮	漫画书	14.00%	mask	8.00%	剧院 帷幕	8.00%
	豹	knot	14.00%	老虎	12.00%	mask	6.00%
	埃及的cat	温莎领带 淋浴帘	10.00% 10.53%	剧院幕布 纱窗	10.00% 7.89%	聚光灯 泰迪熊	8.00% 7.89%
高效 ient Net	虎斑猫	虎斑猫	62.22%	埃及猫	28.89%	波斯猫	4.44%
	安哥拉猫	波斯猫	72.92%	埃及猫	20.83%	虎斑猫	2.08%
	lynx	埃及猫	60.00%	虎斑猫	20.00%	虎斑猫	20.00
	暹罗猫	埃及猫	100.0%	-	-	-	-
	老虎	埃及猫	34.21%	虎斑猫	18.42%	老虎	15.79%
	波斯猫	虎斑猫	100.0%	-	-	-	-
	美洲狮	虎斑猫	100.0%	-	-	-	-
	豹	埃及猫	56.00%	lynx	22.00%	虎斑猫	16.00%
	埃及的cat	漫画书	13.33%	墨西哥无毛犬	13.33%	灯罩	6.67%
		猕猴	33.33%	墨西哥无毛犬	33.33%	靛蓝 彩鹀	33.33%
Conv Next	虎斑猫	虎斑猫	75.00%	埃及猫	15.00%	网站	5.00%
	安哥拉猫	波斯猫	46.67%	埃及猫	35.56	虎斑猫	6.67%
	lynx	虎斑猫	75.00	虎斑猫	25.00%	-	-
	暹罗猫	埃及猫	75.00%	金毛寻回犬	25.00%	-	-
	老虎	虎斑猫	31.82%	埃及猫	31.82%	老虎	9.09%
	波斯猫	暹罗猫	100.0%	-	-	-	-
	美洲狮	网站	100.0%	-	-	-	-
	豹	埃及猫	32.00%	lynx	18.00%	豹	16.00%
	埃及的cat	墨西哥无毛犬	20.83%	mask	12.50%	漫画书	12.50%
		毛皮大衣	50.00%	墨西哥无毛犬	50.00%	-	-