中图分类号：TP3 文献标识码：A 文章编号：1006-8961
论文引用格式：（论文引用格式：）[DOI:]

视觉—语义多模态解纠缠的广义零样本学习

韩阿友[1,2]，杨关[1,2]*，刘小明[1,2]，刘阳[3]

1.中原工学院，计算机学院，河南 郑州 450007；2.河南省网络舆情监测与智能分析重点实验室，河南 郑州 450007；
3.西安电子科技大学，通信工程学院，陕西 西安 710071

摘 要：目的 传统的零样本学习(Zero-Shot Learning，ZSL)旨在依据可见类别的数据和相关辅助信息对训练过程中未见类别的数据进行预测分类，而广义零样本学习(Generalized Zero-Shot Learning，GZSL)中分类的类别既可能属于可见类也可能属于不可见类，这更符合现实的应用场景。基于生成模型的广义零样本学习的原始特征和生成特征不一定编码共享属性所指的语义相关信息，这样导致模型会倾向于可见类，并且分类时忽略了语义信息中与特征相关的有用信息。为了分解出相关的视觉特征和语义信息，提出了视觉—语义多模态解纠缠框架。**方法** 首先使用条件变分自编码器为不可见类生成视觉特征，再通过一个特征解纠缠模块将其分解为语义一致性和语义无关特征。然后，设计了一个语义解纠缠模块将语义信息分解为特征相关和特征无关的语义。其中，利用总相关惩罚来保证分解出来的两个分量之间的独立性，特征解纠缠模块通过关系网络来衡量分解的语义一致性，语义解纠缠模块通过跨模态交叉重构来保证分解的特征相关性。最后使用两个解纠缠模块分离出来的语义一致性特征和特征相关语义信息联合学习一个广义零样本学习分类器。**结果** 实验在四个广义零样本学习公开数据集(AWA2、CUB、SUN 和 FLO)上取得了比 Baseline 更好的结果，在 AwA2 上调和平均值提升了 1.6%，CUB 上提升了 3.2%，SUN 上提升了 6.2%，FLO 上提升了 1.5%。**结论** 在广义零样本学习分类中，本文提出的视觉—语义多模态解纠缠方法经实验证明比基准方法取得了更好的性能，并且优于大多现有的相关方法。

关键词：零样本学习；广义零样本学习；解纠缠表示；变分自编码器；跨模态重构；总相关

Generalized zero-shot learning for visual-semantic multimodal disentangling

Han Ayou[1,2]，Yang Guan[1,2]*，Liu Xiaoming[1,2]，Liu Yang[3]

1.School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China；
2.Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou 450007, China；
3.School of Telecommunications Engineering, Xidian University, Xi'an 710071, China

Abstract: Objective Traditional deep learning models have been widely adopted in many application scenarios and preform effectively, but rely on a large amount of training samples, however, which is very difficult to collect a large amount of training samples in practical applications, bypass the limitation of identify only the classes already present in the training phase (seen classes) and to process the classes which never seen in the training phase (unseen classes) is a challenge, and zero-shot learning provides a good solution to this challenge. Zero-shot learning aims to classify unseen classes for which no training samples are available during the training phase. However, there is a further problem with the complexity in real-world, practically, not only seen classes exist but unseen classes also can be found in real life. Therefore, generalized zero-shot learning, a new approach with the characteristic of more realistic and universality approach has been proposed. Generalized means this method can sample test set from both seen and unseen classes. Existing generalized zero-shot learning methods can be subdivided into two categories, namely embedding-based and generation-based. The former learns a projection or embedding function that

收稿日期:； **修回日期:**
* **通信作者:**杨关 yangguan@zut.edu.cn
基金项目:国家自然科学基金项目(61906141)；陕西省自然科学基金(2020JQ-317)
Supported by:National Natural Science Foundation of China(61906141); Natural Science Foundation of Shaanxi Province,China(2020JQ-317)

associates the visual features of the seen class with the corresponding semantics, while the latter learns a generative model to generate visual features for the unseen class. In previous studies, visual features extracted using the pre-trained deep models (e.g., ResNet101) were not specifically extracted for the generalized zero-shot learning task, and the extracted visual features are not all semantically related to the predefined attributes in dimension, which will lead the model to incline to the seen classes. Most methods ignore useful information related to features present in the semantics when classifying, which has a significant impact on the final classification. In order to disentangle the relevant visual features and semantic information, a generalized zero-shot learning method (MDGZSL) based on the visual-semantic multi-modal disentanglement framework is proposed. **Method** The conditional Variational Auto-Encoders is combined with a disentanglement network, trained in an end-to-end manner, and the proposed disentanglement network is an encoder-decoder structure. The visual features and semantics of the seen classes are first used to train the conditional Variational Auto-Encoders and the disentanglement network. Once the network has converged, the trained generative network is used to generate visual features for the unseen classes, and then the real features of the seen classes and the generative features of the unseen classes are fed into a visual feature disentanglement network to disentangled semantic-consistent and semantic-irrelevant features, followed by a semantic disentanglement network to disentangle the semantic into feature-relevant and feature-irrelevant semantic information. The components disentangled by the two disentanglement networks are fed into the decoder to reconstruct back to the corresponding space, using reconstruction loss to prevent information loss during disentanglement stage. A total correlation penalty module is designed to measure the independence between potential variables disentangled by the disentanglement network. A relational network is designed to maximise the compatibility score between the components disentangled by the visual disentanglement network and the corresponding semantics to learn the semantic consistency of the visual features. The semantic information related to the visual features disentangled by the semantic disentanglement network is fed into the visual disentanglement decoder for cross-modal reconstruction to measure the feature relevance of the semantics. Finally, the semantic consistency features and feature-related semantics disentangled by the two disentanglement networks are jointly learned into a generalized zero-shot classifier for classification. **Result** The proposed method was validated in several experiments on four generalized zero-shot learning open datasets (AwA2, CUB, SUN and FLO), and the proposed method achieved better results than Baseline, with a 3.8% improvement in unseen classes accuracy, a 0.2% improvement in seen classes accuracy and a 1.6% improvement in the harmonic mean on dataset AwA2, the unseen classes accuracy improved by 3.8%, the seen classes accuracy improved by 2.4%, and the harmonic mean improved by 3.2% on dataset CUB, the unseen classes accuracy improved by 10.1%, the seen classes accuracy improved by 4.1%, and the harmonic mean improved by 6.2% on dataset SUN, and the seen classes accuracy improved by 9.1% and the harmonic mean improved by 1.5% on dataset FLO. It was also compared with seven recently proposed generalized zero-shot learning methods (f-CLSWGAN, CANZSL, LisGAN, CADA-VAE, f-VAEGAN-D2, FREE and Cycle-CLSWGAN), with MDGZSL improving the harmonic mean over f-CLSWGAN on the four datasets by 10%, 8.4%, 8.1% and 5.7%. The harmonic mean relative to CANZSL on the four datasets improved by 12.2%, 5.6%, 7.5% and 4.8% respectively. Relative to LisGAN on the four datasets the harmonic mean improved by 8.1%, 6.5%, 7.3% and 3% respectively. Relative to CADA-VAE the harmonic mean over the four datasets improved by 6.5%, 5.7%, 6.9% and 10% respectively. Relative to f-VAEGAN-D2 the harmonic mean over the four datasets improved by 6.9%, 4.5%, 6.2% and 6.7% respectively. Relative to Cycle-CLSWGAN the harmonic mean improved by 5.1%, 8.1% and 6.2% on the CUB, SUN and FLO datasets respectively. Relative to FREE, the harmonic mean on the AwA2, CUB and SUN datasets were improved by 3.3%, 0.4% and 5.8% respectively. The experimental results show that the proposed method achieves better results and thus the effectiveness of the proposed method can be demonstrated. **Conclusion** compared to traditional models; the visual-semantic multimodal disentanglement generalized zero-shot learning (MDGZSL) method proposed by us has the obvious superiority. MDGZSL can disentangled the semantically consistent features in the visual features, the semantic information associated with the features in the semantics, and then learn the disentangled features and semantics jointly a generalized zero-shot learning classifier performs the classification. Moreover, MDGZSL achieves a significant performance improvement by comparing with several related methods on multiple datasets.

Key words: zero-shot learning; generalized zero-shot learning; disentanglement representation; variational auto-encoders;

cross-modal reconstruction; total correlation

0 引　言

随着深度学习的蓬勃发展，许多端到端的深度学习模型已经在很多应用场景上效果斐然。虽然传统的深度学习模型非常成功，但是它们的成功是基于大量的带标记的数据进行训练的。在现实生活中收集大量的标记样本是一个具有挑战性的问题。例如ImageNet(Jia等，2009)是一个大的数据集，它包含1400万张图像，包含21,814个类，但是其中许多类只包含少数图像。此外传统的深度学习模型只能识别训练阶段已有的类别样本，不能处理来自不可见类的样本。这是个非常具有挑战的问题，因为在现实场景中，可能有些类别是没有可训练样本的，比如濒危鸟类等。

人类可以根据先前学习到的经验来学习新的概念，而不必事先看到它们。例如一个人可以很容易地识别出斑马，如果他以前见过马，并且知道斑马看起来像是带有黑白条纹的马(Ji等人，2019)。零样本学习(Zero-Shot Learning，ZSL)(Larochelle等，2008)方法为解决这一挑战提供了一个很好的解决方案。

在零样本学习中训练阶段出现的类别被称为可见类，未出现的类别被称为不可见类。它的目标是训练一个模型，学习语义空间和视觉空间之间的映射。通过语义信息将可见类学习到的知识迁移到不可见类，从而缩小可见类和不可见类之间的差距，然后对不可见类进行分类。但是传统的零样本学习是建立在测试集中只包含不可见类样本的假设之上，这种假设在现实场景中容易被打破。因此出现了一种比零样本学习更有现实意义、更具挑战的广义零样本学习(Generalized Zero-Shot Learning，GZSL)(Chao等，2016)，也就是说，测试集的样本来自可见类和不可见类。

现有的广义零样本学习方法技术主要可以分为两大类：基于嵌入的(Frome等，2013；Liu等，2019；Jiang等，2019；Xian等，2016）和基于生成的（Zhu等，2018；Sanath等，2020；Chen等，2021；Xian等，2018；Keshari等，2020)方法。前者目标是学习一个映射函数，将可见类的视觉特征和其对应的语义向量映射到某一空间中进行后续分类，后者是学习一个生成模型为不可见类生成视觉特征。现在大多数的广义零样本学习方法是利用在ImageNet上预训练的深度模型来提取相应的视觉特征，比如ResNet101(He等，2016)和VGG-19(Simonyan等，2014)等。然而现有的大多广义零样本学习方法中忽略了语义和视觉的相关性。因为在模型学习的过程中，原始特征和生成特征在维度上并不是都与预定义属性在语义上相关，这将导致视觉在维度上与语义产生偏见，并导致了对不可见类的负迁移，如图1所示，红色方框标注的“耳朵”维度与注释属性在语义上无关，从这些语义无关的视觉特征中学习可能会影响模型对不可见类的泛化。此外大多数方法在分类过程中都忽略了丰富的语义信息，并且在语义信息中也存在着与分类无关、特征无关的信息。这将会影响分类结果，如图2所示，红色划线部分语义是和视觉特征无关的，比如在注释属性中存在的“Ocean”语义对于“猫”的视觉特征是无关的，并且对于最终的视觉—语义联合分类也会产生不好的影响。

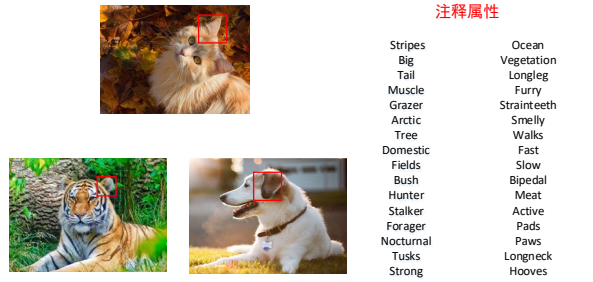


图 1　语义无关的视觉特征图示

Fig.1 The illustration of visual features (red boxes) that are not associated with the annotated attributes

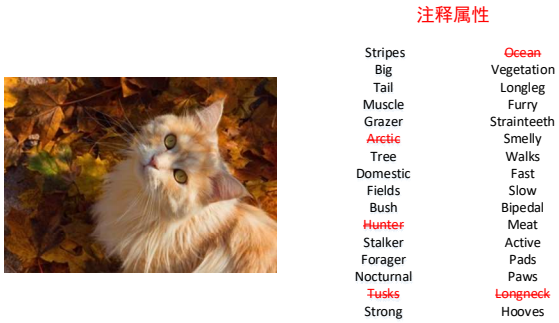


图 2　特征无关的语义注释属性图示

Fig.2 The illustration of annotated attributes (red lines) that are not associate with the visual features

为了解决上述问题，本文提出了视觉—语义多模态解纠缠框架(Multimodal Disentangling Generlized Zero-Shot Learning，MDGZSL)。通过视觉—语义解纠缠框架来提取出语义一致性特征和特征相关的语义信息，设计了一个总相关惩罚结构

衡量分解的潜在变量之间的独立性，设计一个语义一致性衡量网络来衡量分解出来的视觉特征的语义一致性。然后将视觉特征分解的潜层输出和语义信息分解的潜层输出进行跨模态交叉重构。对于视觉特征交叉重构输入的是语义信息分解输出中与特征相关的信息，并使用该操作来指导语义解纠缠框架分解出与特征相关的语义信息。最后将语义一致性特征和特征相关的语义信息联合学习一个广义零样本学习分类器。在四个公开数据集上的实验结果验证了本文提出方法的有效性。

本文的贡献如下：

1）提出发现了在提取的视觉特征中并不是都与预定属性在语义上相关，这将会导致产生语义偏见，并且在语义信息中存在与分类无关以及特征无关的冗余信息。

2）提出一个视觉—语义解纠缠框架，用来提取出视觉特征中语义一致性特征和预定义属性中特征相关的语义信息。设计了一个跨模态交叉重构模块来指导语义解纠缠能更好的分解出与特征相关的语义信息，采用关系网络来学习视觉解纠缠分解出语义一致性表示。最后将解纠缠模块分解后的特征和语义联合学习一个广义零样本学习分类器进行分类。

3）在广义零样本学习四个公开数据集上进行多次实验，通过解纠缠框架学习到的语义一致性视觉特征和特征相关的语义信息能够提高分类性能并优于对比的基准方法，证明了所提视觉—语义解纠缠思想的有效性。

1 相关工作

1.1 广义零样本学习

广义零样本学习是比零样本学习更有现实意义、更具挑战的情况，即就是在测试集样本中既有可见类也有不可见类。由于在训练阶段不可见类的视觉样本不可用，这导致经验风险最小化变得不可靠(Wang等，2020)。为了克服这些限制，利用语义作为不可见类的中间表示，这种语义通常是手动定义的属性(Lampert等，2013)。

广义零样本学习的方法有：f-CLSWGAN(Xian等，2018)中利用Wasserstein GANs(WGAN)(Arjovsky等，2017)来合成逼真的视觉特征。CADA-VAE(Schonfeld等，2019)利用两个对齐的变分自动编码器来学习不同模式之间的共享潜在表示。TF-VAEGAN(Sanath等，2020)将VAE和GAN相结合来生成视觉特征，再通过一个语义解码器将视觉特征解码出语义属性，并且提出了一个反馈模块，将语义解码器的潜层表示作为输入反馈回生成器的潜层来提高生成特征的质量。CANZSL(Chen等，2020)提出了周期一致对抗网络，首先从有噪声的文本中合成视觉特征，然后采用逆对抗网络将生成特征转换为文本，以确保合成的视觉特征能够准确地反映语义表征。OT-ZSL(Wang等，2021)利用一个条件生成模型从可见类属性生成可见类特征，并在生成特征分布和真实特征分布之间建立最优传输，利用基于属性的正则化器对生成模型和最优传输进行迭代优化，进一步增强了所生成特征的鉴别能力。FREE(Chen等，2021)指出在ImageNet上训练的特征提取器忽略了ImageNet和GZSL数据集之间的偏差，这种偏差会导致广义零样本学习任务的视觉特征质量低下，因此提出了一种特征细化的方法，采用一种自适应边缘中心损失，它与语义循环一致性损失相结合，引导特征细化模块学习类和语义相关的表示。Zhao等人(2021)提出了一种基于子空间学习和重构的零样本分类方法来解决知识迁移过程中的信息损失和域偏移问题。在人体行为识别领域中Lyu等人(2021)为了研究多种模态数据对零样本人体动作识别的影响，提出了一种基于多模态融合的零样本人体动作识别ZSAR-MF框架，该框架能有效地融合传感器特征和视频特征。

1.2 生成模型

最近的广义零样本学习中使用生成模型的方法取得了先进的性能。生成模型可以为不可见类合成大量的视觉特征，一旦为不可见类生成了视觉特征，那么零样本学习的问题就变成了一个相对简单的监督分类问题。两种常用的生成模型是生成对抗网络(Generative Adversarial Networks，GAN)(Goodfellow等，2014)和变分自编码器(Variational Auto-Encoder，VAE)(Kingma等, 2014)，这两种模型在基于生成方法的广义零样本学习任务中被广泛使用。其中Xian等人(2018)设计了一个带有分类损失的条件WGAN模型称为f-CLSWGAN，将语义特征集成到生成器和鉴别器中。SP-GAN(Ma等, 2020)设计了一种保留相似性损失和分类损失。SR-GAN(Ye等, 2019)使用语义矫正网络来矫正特征。CVAE-ZSL(Mishra等, 2018)采用神经网络对编码器进行建模，SE-GZSL(Verma等, 2018)设计了一个循环一致性损失函数，配备了鉴别器驱动的反馈机制，

将真实样本或生成的样本映射回相应的语义表示。在这些方法中，使用的原始特征和生成特征中存在着与预定义属性不相关的特征，并且在分类阶段要么忽略了丰富的语义信息，要么在使用语义信息时没有解耦出语义信息中与特征相关的部分。如何让模型提取它们是个关键。

1.3　解纠缠表示

解纠缠指的是一种表示特征之间的独立性。总相关性(Total Correlation，TC)(Kim等，2018)是对多个随机变量独立性的测量。在信息论中，总相关是互信息对随机变量的许多推广之一。它是最近解纠缠方法的一个关键组成部分。FactorVAE(Kim等，2018)建议使表示的分布进行阶乘来分离特征，从而实现跨维度的独立性。Higgins等人(2017)提出的beta-VAE是一种无监督的视觉解纠缠表示学习方法，通过调整KL项的权重来平衡解纠缠因子的独立性和重构性能。Chen等人(2016)提出的InfoGAN通过最大化潜层变量与原始变量之间的互信息来实现解纠缠。LFZRL(Tong等，2019)提出了一种分层分解方法来学习有区别的潜在特征。

2　问题定义

在零样本学习中，数据集类别分为可见类s和不可见类u，标签分别为y_s和y_u，$y_s \cap y_u = \varnothing$。假设训练数据集$D_s^{tr} = \{(x_s, a_s, y_s)\}$，仅由可见类中标记的样本组成，其中$x_s \in X^s$表示可见类视觉特征，$a_s \in A^s$是可见类相关的语义描述符(比如语义属性)，$y_s \in Y^s$表示可见类的类标签。测试集$D_u^{te} = \{x_u, a_u, y_u\}$，其中在训练期间不可见类的视觉特征$x_u$不可用。传统的零样本学习旨在学习测试集$D^{te} = \{x_u\}$上评估的分类器$f^{ZSL}: X^u \to Y^u$。然而在广义零样本学习中，测试集$D^{te}$由可见类和不可见类共同组成，即就是学习在所有的特征上评估的分类器$f^{GZSL}: X \to Y^u \cup Y^s$，本文主要研究的是广义零样本学习的分类问题。

3　方法

为了同时得到语义一致性的视觉特征和特征相关的语义属性，本文提出了一种基于总相关惩罚的视觉—语义解纠缠框架。分解的视觉特征通过一个关系网络来保证语义一致性，分解的语义信息通过视觉交叉重构来保证特征相关性。最后通过语义一致性视觉特征和特征相关的语义信息结合进行广义零样本学习的分类。

3.1　模型架构

所提模型架构如图3所示，主要由条件变分自编码器、视觉—语义解纠缠模块、语义一致性特征衡量网络、总相关惩罚和视觉—语义跨模态重构组成。框架的视觉输入是由预训练的ResNet101(He等，2016)提取的图像特征，语义输入是人工定义的属性。

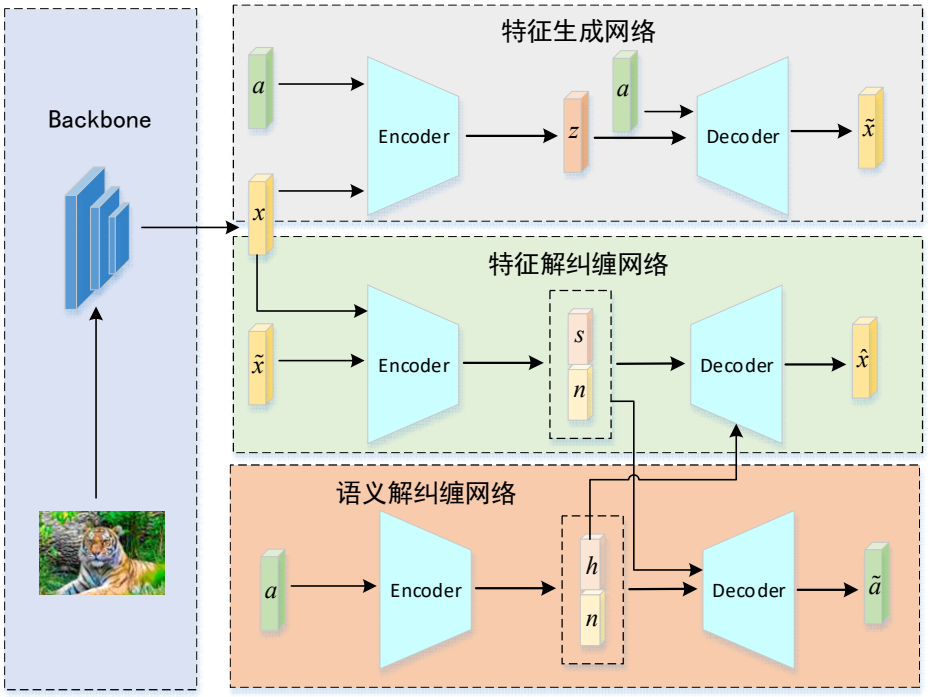


图 3　模型架构

Fig.3 Model architecture

3.2 视觉特征生成

为了通过语义信息来生成视觉特征，这里使用条件变分自编码器(cVAE)(Sohn等，2015)为不可见类生成相应的视觉特征。cVAE学习数据和潜在表示的分布之间的关系。它由编码器以及解码器组成。其中编码器将特征空间映射到潜在空间，解码器将潜在空间映射回特征空间，它们分别将类的描述符作为条件。cVAE的目标函数可以写成：

$$L_{cVAE} = -KL\left[q(z|x,a)\|p(z|a)\right] + \mathrm{E}_{q(z|x,a)}\left[\log p(x|z,a)\right] \tag{1}$$

其中x表示视觉特征，a表示语义信息，z表示编码器生成的潜在变量。式中第一项为$q(z|x,a)$和$p(z|a)$两个分布之间的KL散度，约束编码器匹配分解后的先验分布，例如高斯分布。第二项为重构特征和原始特征之间的重构误差。

为了获得变分下界的可微估计量，使用了一种被称为重参数化的技巧：

$$z = \mu(x) + \sigma(x) \odot \varepsilon \tag{2}$$

其中$\mu(x)$和$\sigma(x)$是编码器的输出，代表后验分布的均值和方差。$\odot$表示两个张量对应元素的乘积，$\varepsilon \sim \mathrm{N}(0,1)$是一个辅助噪声变量。

3.3 解纠缠模块

对于广义零样本学习的数据集来说，通过预训练的深层模型(比如ResNet101)提取的视觉特征并不完美，因为视觉特征并不是在所有的维度上都和预定义属性在语义上相关。在这些数据集中，类别通常是相关的(比如CUB数据集都对应于鸟类)，因此提取的特征可能包含冗余信息。这里将视觉特征分解成语义一致性特征s和语义无关特征n，使用一个关系网络来衡量语义一致性。在语义信息中也存在着跟特征相关的信息，将语义信息分解成特征相关语义h和特征无关语义n，使用视觉特征交叉重构来保证分解的语义和视觉特征相关。为了加强分解后两个分量的独立性，使用总相关性来衡量。

3.4 视觉—语义跨模态重构

解纠缠模型使用和cVAE相同的编码器—解码器结构。分别对两个解纠缠结构的解码器输出特征和编码器输入特征计算重构损失。在视觉和语义两个解纠缠中，提出了一个跨模态重构损失，把视觉解纠缠的编码器E_1的输出送入到语义解纠缠的解码器D_2中来重构语义信息，然后把语义解纠缠的编码器E_2的输出中与特征相关的分量送入到视觉解纠缠的解码器D_1中来重构视觉特征。两个解纠缠模块的重构损失分别为：

$$L_{visual} = \sum_{x \in X^s} \left\| x - D_1(s,n) \right\|^2 \tag{3}$$

$$L_{semanic} = \sum_{a \in A^s} \left\| a - D_2(h,n) \right\|^2 \tag{4}$$

视觉—语义跨模态交叉重构损失为：

$$L_{semantic-visual} = \sum_{x \in X^s} \left\| x - D_1\left[(s,n),h\right] \right\|^2 \tag{5}$$

$$L_{visual-semantic} = \sum_{a \in A^s} \left\| a - D_2\left[(h,n),(s,n)\right] \right\|^2 \tag{6}$$

其中对于特征解纠缠编码器输出s和n中，s是语义一致性特征，n是语义无关特征。对于语义解纠缠编码器输出h和n中，h是特征相关的语义向量，n是特征无关的语义向量。解纠缠模块总的重构损失为：

$$L_{rec} = L_{semantic-visual} + L_{visual-semantic} + L_{visual} + L_{semanic} \tag{7}$$

其中，使用均方误差MSE来计算原始视觉特征和重构视觉特征、原始语义向量和重构语义向量之间的重构损失。

3.5 语义一致性特征

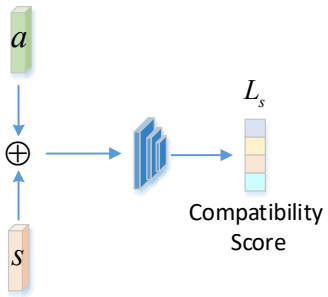


图 4 语义一致性衡量网络

Fig.4 Semantic consistency measurement network

在视觉解纠缠框架中采用一个关系网络(RelationNet，RN)(Sung等，2018)来作为语义一致性衡量网络，最大化视觉特征解纠缠模块分解的潜在表示s和相对应的语义a之间的相容性得分(Compatibility Score，CS)来学习语义一致性特征s。语义一致性衡量网络如图4所示。关系网络RN学习潜在表示和对应语义向量之间的成对关系。RN的输入是潜在表示s和对应唯一语义向量a组成的对。组成的对如果匹配成功CS值为1，不匹配为CS为0，结构表示为：

$$CS\left(s_{(t)},a_{(t)}\right) = \begin{cases} 0 & if\ y_{(t)} \neq y_{(c)} \\ 1 & if\ y_{(t)} = y_{(c)} \end{cases} \tag{8}$$

其中t和c表示一个$batch$中第t个语义一致性表示和第c个唯一语义向量，$y_{(t)}$和$y_{(t)}$表示$s_{(t)}$和$a_{(c)}$的类标签。

利用等式(8)中的定义的CS，使用带有Sigmoid激活函数的关系网络为每一对(s, a)学习一个0到1的相容性得分。然后使用以下损失函数来优化s：

$$L_s = \sum_{t=1}^{B}\sum_{c=1}^{N}\left\| RN\left(s_{(t)}, a_{(t)}\right) - CS\left(s_{(t)}, a_{(c)}\right)\right\|^2 \quad (9)$$

其中B为$batch\ size$，N为一个$batch$中唯一语义向量的数量。使用均方误差来优化该损失，来保证视觉解纠缠分解出语义一致性特征。

3.6　总相关惩罚

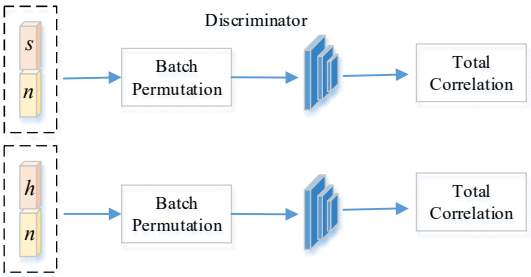


图 5　鉴别器

Fig.5 Discriminator

为了促进视觉特征和语义信息解纠缠模块能更好的分离特征和语义，这里设计了一个总相关惩罚来鼓励视觉解纠缠模块分离出的语义一致性和语义无关特征之间的独立性，语义解纠缠模块分离出特征相关和特征无关语义之间的独立性。这里使用语义解纠缠来对总相关惩罚进行展开解释。对于语义信息的解纠缠分解出的两个分量，可看作是独立的，并且来自不同的条件分布，语义向量的潜在分量的分布分别为：

$$h \sim \Phi_1(h \mid a), n \sim \Phi_2(n \mid a) \quad (10)$$

其中h为语义解纠缠模块分解出的与特征相关的语义信息，n为分解出的与特征无关的语义信息，a为需要被分解的语义。总相关性可表示为：

$$TC = (\Phi \| \Phi_1 \cdot \Phi_2) \quad (11)$$

其中$\Phi : \Phi(h, n \mid a)$是语义解纠缠中分解出两个分量的联合条件概率。为了更好的逼近总相关性，使用密度比估计以对抗的方式区分两个分布中的样本(Chen等，2021)，使用一个鉴别器$Dis(t)$的输出估计独立分量的概率，鉴别器模型如图5所示。

$$TC = E_{\Phi}(\log\frac{\Phi}{\Phi_1 \cdot \Phi_2}) \approx E_{\Phi}(\log\frac{Dis(t)}{1 - Dis(t)}) \quad (12)$$

其中t=[h, n]。鉴别器损失为：

$$L_{dis} = \log Dis(t) + \log(1 - Dis(\tilde{t})) \quad (13)$$

其中$\tilde{t}$是在$batch$维度上随机打乱每个h和n然后再拼接得到的结果。语义向量分解得到独立的分量后，然后将此分量送入到视觉特征交叉重构中得到重构特征，最小化该视觉交叉重构损失来鼓励语义解纠缠模块分离出与视觉特征相关的语义向量。

3.7　模型算法

所提方法MDGZSL训练算法如下所示：

算法 MDGZSL 训练算法

输入：可见类视觉特征X^s、语义向量A^s及其标签Y^s

输出：训练好的生成网络和解纠缠网络

While 模型不收敛：
　随机选择一个批次数据$\{x_{(t)}^s, y_{(t)}^s\}_{t=1}^{B}, \{a_{(c)}^s\}_{c=1}^{N}$
　FOR iters = 0 to n
　　根据式(1)计算条件自编码器损失 L_{cVAE}
　　根据式(3)和(4)计算视觉和语义解纠缠模块重构损失L_{visual}和$L_{semantic}$
　　根据式(5)和(6)计算解纠缠模块中视觉和语义跨模态重构$L_{semantic-v}$　和$L_{visual-seman}$
　　根据式(7)计算总的重构损失L_{rec}
　　根据式(13)计算鉴别器损失$L_{dis}^{visual}$and $L_{dis}^{semantic}$逼近视觉和语义解纠缠总相关惩罚
　　反向传播 $\lambda_3(L_{dis}^{visual} + L_{dis}^{semantic})$
　　根据式(1),(7)和(9)计算总损失$L_{all1} = L_{cVAE} + L_{rec} + \lambda_1 L_s$
　　反向传播 L_{all}
　END FOR
　随机选择一个批次数据$\{x_{(t)}^s, y_{(t)}^s\}_{t=1}^{B}, \{a_{(c)}^s\}_{c=1}^{N}$
　根据式(1),(7),(9)和(12)计算总损失$L_{all2} = L_{cVAE} + L_{rec} + \lambda_1 L_s + \lambda_2 TC$
　反向传播 L_{all2}
END While

4　实验

为了验证本文所提出的方法对广义零样本学习的图像分类任务的有效性，在四个公开数据集上进行实验。从参数分析、消融实验两个方面进行实验分析并展示实验效果。

4.1　实验数据

实验使用四个基准数据集来评估所提模型的

性能，分别是：Animals with Attributes2(AwA2)(Lampert等，2013)，Caltech-UCSD Birds-200-2011(CUB)(Wah等，2011)，Oxford Flowers(FLO)(Nilsback等，2008)和SUN Attribute(SUN)(Patterson等，2012)。其中CUB数据集中包含200种鸟类，其中150种是可见类，50种是不可见类，每个类别有312个属性的注释。AwA2数据集是常用于动物分类，由40个可见类和10个不可见类组成，每个类别都有85个属性的注释。SUN是一个大型场景风格的数据集，包含645个可见类和72个不可见类，每个类别有102个属性的注释。FLO数据集包含102个花卉类别，82个可见类和20个不可见类，注释属性有1024维。表1列出了每个数据集的详细信息。

表1 数据集统计

Table 1 Statistics of datasets

数据集	视觉特征维度	语义特征维度	可见类别数	不可见类别数
CUB	2048	312	150	50
AwA2	2048	85	40	10
SUN	2048	102	645	72
FLO	2048	1024	82	20

4.2 评估方法

在广义零样本学习任务上评估精度使用的是调和平均值，它计算的是可见类和不可见类的联合精度，公式为：

$$H=\frac{2\times U\times S}{U+S} \tag{14}$$

其中U表示在不可见类图像上每个类别的平均精度，衡量不可见类样本的分类能力。S表示在可见类图像上每个类别的平均精度，衡量可见类样本的分类能力。H是调和平均值，来衡量GZSL任务的性能。

4.3 实验设置

按照大部分方法的设置，首先利用预训练的ResNet101来提取维度为2048的图像特征。语义特征是由人工注释的每个类别的描述。cVAE和解纠缠模块的编码器和解码器都是由多层感知机(Multilayer Perceptron，MLP)组成。在cVAE中，隐藏层维度为2048，生成的潜在表示维度为20。解纠缠模块的隐藏层维度是可以调节的参数。关系网络的隐藏层维度2048。鉴别器模块是带有Sigmoid 激活函数的单层感知机。

所提方法是由PyTorch实现，并采用Adam(Kingma等，2015)优化器进行优化。学习率是一个可调节的超参数，批次大小设置为64。当cVAE和解纠缠训练好后，使用cVAE中的生成器来为不可见类生成大量样本。之后将可见类的训练特征和不可见类的生成特征送入到视觉解纠缠模块中提取语义一致性特征，把语义信息送入语义解纠缠模块提取特征相关的语义向量。最后将分解出来的语义一致性特征和特征相关语义信息联合共同学习一个广义零样本学习分类器，然后计算相应的指标。其中分类过程如图6所示。

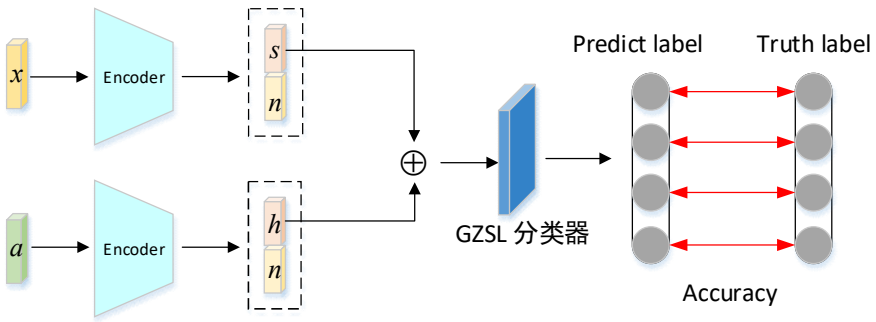


图 6 分类图

Fig.6 Classifier architecture

4.4 对比相关方法

为了证明所提方法的有效性，选择了7种不同的方法进行实验对比。对比方法如下：

f-CLSWGAN(Xian等，2018)提出了一种使用GAN在特征空间上生成数据，并添加一个辅助分类器来提高生成器性能的方法来解决ZSL问题，相比于直接生成图像，该方法能取得更好的性能。

CANZSL(Chen等，2020)提出了一个基于自然语言语义空间的循环一致性对抗网络。该网络使用的是带有不相关词的自然语言来生成视觉特征而

不是使用以往人工注释的语义信息，然后由语义特征生成器将合成的视觉特征反译回相应的语义空间。

Cycle-CLSWGAN(Felix 等， 2018) 提出使用cycle-consistent loss作为正则化项来训练GAN，使得所生成的视觉特征能够重构它的原始语义特征。解决了基于生成方法的GZSL模型中存在从语义到视觉特征的生成过程没有约束的问题。

RREE(Chen等，2021)提出一种自适应边缘中心损失，它和语义循环一致性损失相结合对视觉特征进行细化，减轻了ImageNet和GZSL基准数据集之间的跨数据集偏差。

LisGAN(Li等，2019)提出了在GAN生成器中引入灵魂样本正则化方法来解决视觉对象的多视图质量问题，并在分类阶段提出使用级联分类器来微调精度。

CADA-VAE(Schonfeld等，2019)使用VAE对视觉特征和类别描述进行编码解码，对这两个模态进行对齐，在隐空间中使用这两个模态共同构建分类器。

f-VAEGAN-D2(Xian等，2019)提出了一个直推式特征生成网络，使用VAE和WGAN结合进行信息互补来生成更鲁棒的特征，额外使用一个非条件的鉴别器来学习不可见类的流形。

4.5　实验结果

在四个公开数据集上与Baseline和其它七个方法进行多次实验对比。表2展示了MDGZSL与其它相关方法之间的对比实验结果。除了FLO数据集外，所提的方法都优于所有比较的方法，并且在四个数据集上比Baseline方法结果更好。

MDGZSL相对于CANZSL(Chen等，2020)在数据集AwA2上的U、S和H分别提高了17.6%，3.6%和12.2%，CUB上的U、S和H分别提高了7.4%，3%和5.6%，SUN上的U、S和H分别提高了11.5%，5.2%和7.5%，FLO上的U、S和H分别提高了1.5%，10.8%和4.8%。因为CANZSL并没有将视觉特征中存在的语义无关特征进行分离，这会产生语义偏见，并且在分类时该方法使用的是一个k-最近邻算法进行分类，这也忽略了语义信息对分类结果的影响。因此MDGZSL方法比之性能更好。

其中相对于参照的Baseline方法，所提方法MDGZSL在数据集AwA2上U提高了2.7%，S提高了0.2%，H提高了1.6%，CUB上U提高了3.8%，S提高了2.4%，H提高了3.2%，SUN上U提高了10.1%，S提高了4.1%，H提高了6.2%，FLO上S提高了9.1%，H提高了1.5%。由于Baseline只考虑了对视觉特征的解纠缠，忽略了在语义信息中与分类特征相关的信息，而MDGZSL正是发现了这个问题并解决了这个不足，从而提高了Baseline的性能。

从实验结果看，所提的视觉—语义解纠缠方法可以学习到视觉空间中语义一致性特征和语义空间中特征相关的语义，并且能够提高广义零样本学习分类性能。由此结果可以证明所提方法的有效性。

表 2 方法结果对比(%)

Table 2 Comparison of method results (%)

方法	AwA2			CUB			SUN			FLO		
	U	S	H	U	S	H	U	S	H	U	S	H
f-CLSWGAN	56.1	65.5	60.4	43.7	57.7	49.7	42.6	36.6	39.4	59.0	73.8	65.6
CANZSL	49.7	70.2	58.2	47.9	58.1	52.5	46.8	35.0	40.0	58.2	77.6	66.5
LisGAN	52.6	**76.3**	62.3	46.5	57.9	51.6	42.9	37.8	40.2	57.7	83.8	68.3
CADA-VAE	55.8	75.0	63.9	51.6	53.5	52.4	47.2	35.7	40.6	51.6	75.6	61.3
f-VAEGAN-D2	57.6	70.6	63.5	48.4	60.1	53.6	45.1	38.0	41.3	56.8	74.9	64.6
FREE	60.4	75.4	67.1	55.7	59.9	57.7	47.4	37.2	41.7	**67.4**	84.5	**75.0**
Cycle-CLSWGAN	-	-	-	**59.3**	47.9	53.0	33.8	**47.2**	39.4	59.2	72.5	65.1
SDGZSL(Baseline)	64.6	73.6	68.8	51.5	58.7	54.9	48.2	36.1	41.3	62.2	79.3	69.8
MDGZSL(Ours)	**67.3**	73.8	**70.4**	55.3	**61.1**	**58.1**	**58.3**	40.2	**47.5**	59.7	**88.4**	71.3

注：加粗字体为每列最优值。

4.6　实验分析

在本小节中，我们从参数分析和消融实验两个方面对提出的方法进行实验分析。

4.6.1　参数分析

为了讨论参数对解纠缠模块的影响，这里选择在AwA2 数据集上对解纠缠模块中关系网络损失

L_s的权重λ_1，总相关惩罚损失TC的权重λ_2和鉴别器损失L_{dis}的权重λ_3设置不同的值进行多次实验。如图7所示，这里展示了三个参数对广义零样本学习性能的影响。其中图7(a)是设置其它两个参数固定为λ_2= 1.0, λ_3= 0.5的情况下，关系网络权重λ_1对广义零样本学习性能影响图，在λ_1= 0.7时性能最好。图7(b)表示固定λ_1= 0.7, λ_3= 0.5情况下，总相关惩罚的权重λ_2对广义零样本学习性能影响图，在λ_2=0.9时性能最好。图7(c)表示固定λ_1= 0.7，λ_2= 0.9的情况下，鉴别器损失的权重λ_3的值对GZSL性能影响图，在λ_3= 0.5 时性能最好。综合这三组参数分析实验，找到了在AwA2 数据集上三个参数最优的组合。即就是当λ_1= 0.7, λ_2= 0.9, λ_3= 0.5的情况下广义零样本学习性能达到了最优，调和平均值超过了所有的对比方法。

4.6.2 消融实验

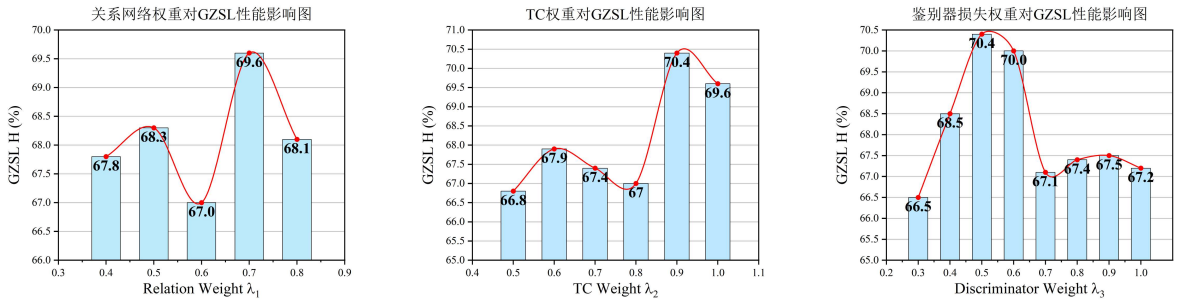
为了验证视觉特征和语义信息联合分类、语义解纠缠和跨模态重构的有效性，分别在四个数据集上进行了消融实验。消融实验结果如表3所示。

其中第一个消融实验是在Baseline中将分离的语义一致性特征和未分离的语义信息联合学习一个分类器得到的结果，在CUB和SUN上的调和平均值H都得到了提升，AwA2和FLO上的S指标都得到提高，调和平均值也取得了相当的结果，这表示加入语义信息进行联合学习对分类有着很大的作用。

第二个消融实验是在Baseline中只添加语义解纠缠模块，这只是单纯的分离出两个独立的语义分量，并没有对分离出来的分量施加约束。可以发现在SUN数据集上性能得到了提升，但是在其它数据集上效果变差，这是因为没有对分解出的语义信息施加约束。

第三个消融实验在加入语义解纠缠模块后的基础上再添加跨模态重构损失，即就是本文所提方法。可以看出在四个数据集上几乎所有评估指标都取得了显著的提升，因为添加的损失会对语义解纠缠模块进行指导，约束语义解纠缠模块分离出特征相关和特征不相关的两个独立的分量，有了丰富的特征相关的语义向量后，将语义一致性特征和特征相关的语义信息联合学习广义零样本学习分类器。

本小节的消融实验进一步证实所提方法能够提高Baseline在多个数据集上的性能并且优于大多数相关方法的性能，更加充分地证明了方法的有效性。



(a) RelationNet weight λ_1 (b) TC weight λ_2 (c) Discriminator weight λ_3

图 7 参数分析

Fig.7 Parameters analysis

((a)RelationNet weight λ_1； (b) TC weight λ_2； (c)Discriminator weight λ_3)

表 3 消融实验(%)

Table 3 Ablation experiments (%)

方法	AwA2			CUB			SUN			FLO		
	U	S	H	U	S	H	U	S	H	U	S	H
Baseline	64.6	73.6	68.8	51.5	58.7	54.9	48.2	36.1	41.3	**62.2**	79.3	69.8
+Attribute	60.8	**77.5**	68.2	**56.8**	60.0	**58.3**	44.4	40.9	42.6	58.9	81.4	68.4
+Attribute Distentangle	60.2	74.7	66.7	54.4	53.5	54.0	46.3	**42.3**	44.2	54.0	79.4	64.3
+Attribute Disentangle+Cross Modal Loss (Ours)	**67.3**	73.8	**70.4**	55.3	**61.1**	58.1	**57.9**	40.2	**47.4**	59.7	**88.4**	**71.3**

注：加粗字体为每列最优值。

5 结论

在使用解纠缠表示的零样本学习方法中，往往都忽略了语义信息。因此本文提出了视觉—语义多模态解纠缠的广义零样本学习分类方法。具体而言，从视觉特征中分解出语义一致性特征和语义无关特征，从人工注释的类别描述（语义属性）中进一步分解出特征相关和特征无关的语义信息。本文设计了一个总相关惩罚来鼓励两个解纠缠框架分离出来的潜在变量之间的独立性，采用关系网络来衡量分解出视觉特征的语义一致性，设计一个跨模态交叉重构的方式来保证分解出来的语义信息是与特征相关的。最后将分解出来的特征相关语义分量和语义一致性视觉特征分量相结合训练一个分类器进行广义零样本学习分类。将解纠缠模块与条件变分自编码器结合，以端到端的方式进行训练。在四个公开数据集上评估所提方法。大量实验证明，我们提出的方法取得了比基准模型更好的效果并优于大多相关方法。

参考文献（References）

Arjovsky, Martin, Soumith Chintala, and Léon Bottou. 2017.Wasserstein generative adversarial networks//Proceedings of the 34th International Conference on Machine Learning.Sydney, NSW, Australia: ICML.org: 214-223[DOI: 10.5555/3305381.3305404]

Chao Wei-Lun, Soravit Changpinyo, Gong Boqing and Sha Fei. 2016.An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild//Bastian Leibe, Jiri Matas, Nicu Sebe, Max Welling.eds.European Conference on Computer Vision. Amsterdam: Springer: 52-68[DOI: 10.1007/978-3-319-46475-6_4]

Chen Shiming, Wang Wenjie, Xia Beihao, Peng Qinmu, You Xinge, Zheng Feng and Shao Ling.2021. FREE: Feature Refinement for Generalized Zero-Shot Learning//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision.Montreal, QC, Canada: IEEE: 122-131[DOI: 10.1109/ICCV48922.2021.00019]

Chen Xi, Duan Yan, Houthooft Rein, Schulman John, Sutskever Ilya and Abbeel Pieter.2016.InfoGAN: interpretable representation learning by information maximizing generative adversarial nets//Proceedings of the 30th International Conference on Neural Information Processing Systems.Barcelona, Spain: Curran Associates Inc: 2180-22188[DOI: 10.5555/3157096.3157340]

Chen Zhi, Li Jingjing, Luo Yadan, Huang Zi and Yangyang. 2020.CANZSL: Cycle-Consistent Adversarial Networks for Zero-Shot Learning from Natural Language//Proceedings of the 2020 IEEE Winter Conference on Applicationsof Computer Vision.Snowmass, CO, USA: IEEE: 863-872[DOI: 10.1109/WACV45572.2020.9093610]

Chen Zhi, Luo Yadan, Qiu Ruihong, Wang Sen, Huang Zi, Li Jingjing and Zhang Zheng.2021.Semantics Disentangling for Generalized Zero-Shot Learning//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision.Montreal: IEEE:8692-8700 [DOI: 10.1109/ICCV48922.2021.00859]

Felix Rafael, B.G.Vijay Kumar, Ian Reid and Gustavo Carneiro. 2018.Multi-modal Cycle-Consistent Generalized Zero-Shot Learning//Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. eds.European Conference on Computer Vision.Munich, Germany:Springer:21-37[DOI: 10.1007/978-3-030-01231-1_2]

Frome, Andrea and Corrado, Greg S and Shlens, Jon and Bengio, Samy and Dean, Jeff and Ranzato, Marc'Aurelio and Mikolov, Tomas.2013.Devise: A deep visual-semantic embedding model.Advances in neural information processing systems, 2:2121-2129[DOI: 10.5555/2999792.2999849]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozairy S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press:2672-2680

He Kaiming, Zhang Xiangyu, Ren Shaoqing and Sun Jian.2016. Deep Residual Learning for Image Recognition//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition.Las Vegas, NV, USA: IEEE: 770-778[DOI: 10.1109/CVPR.2016.90]

Higgins Irina, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matt Botvinick, Shakir Mohamed, Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework //Proceedings of the 5th International Conference on Learning Representations. Toulon, France:ICLR

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei.2009.ImageNet: A large-scale hierarchical image database//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition.Miami: IEEE:248-255 [DOI: 10.1109/CVPR.2009.5206848]

Jiang Huajie, Wang Ruiping, Shan Shiguang and Chen Xilin.2019.Transferable Contrastive Network for Generalized Zero-Shot Learning//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision.Seoul, Korea: IEEE: 9764-9773[DOI: 10.1109/ICCV.2019.00986]

Ji Z, Wang H R, Yu Y L and Pang Y W. 2019. A decadal survey of zero-shot image classification (in Chinese). Journal of Sci Sin Inform, 49(10):1299-1320(冀中, 汪浩然, 于云龙, 庞彦伟. 2019. 零样本图像分类综述: 十年进展. 中国科学: 信息科学，49(10):1299-1320)[DOI:10.1360/N112018-00312]

Keshari, Rohit and Singh, Richa and Vatsa, Mayank.2020.Generalized Zero-Shot Learning via Over-Complete Distribution//Proceedings of the 2020 IEEE/CVF International Conference on Computer Vision and Pattern Recognition.Seattle, WA, USA: IEEE:13297-13305[DOI: 10.1109/CVPR42600.2020.01331]

Kim, Hyunjik and Mnih, Andriy. 2018. Disentangling by factorising//Proceedings of the 35th International Conference on Machine Learning. PMLR: 2649-2658

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR

Kingma D P and Welling M. 2014. Auto-encoding variational Bayes//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR

Lampert, Christoph H, Hannes Nickisch and Stefan Harmeling.2013.Attribute-Based Classification for Zero-Shot Visual Object Categorization.IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(3):453-465[DOI: 10.1109/TPAMI.2013.140]

Larochelle, Hugo and Erhan, Dumitru and Bengio, Yoshua.2008. Zero-Data Learning of New Tasks//Proceedings of the 23rd national conference on Artificial intelligence. Chicago, Illinois:AAAI Press:646-651[DOI: 10.5555/1620163.1620172]

Li Jingjing, Jing Mengmeng, Lu Ke, Ding Zhengming, Zhu Lei and Huang Zi.2019.Leveraging the Invariant Side of Generative Zero-Shot Learning//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Long Beach, CA, USA: IEEE: 7394-7403[DOI: 10.1109/CVPR.2019.00758]

Liu Shichen, Long Mingsheng, Wang Jianmin, Jordan and Michael I.2018.Generalized zero-shot learning with deep calibration network. Advances in neural information processing systems, 2009-2019[DOI: 10.5555/3326943.3327129]

Liu Yang, Guo Jishun, Cai Deng and He Xiaofei.2019.Attribute Attention for Semantic Disambiguation in Zero-Shot Learning//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision.Seoul, Korea: IEEE: 6697-6706[DOI: 10.1109/ICCV.2019.00680]

Lyu L L, Huang Y, Gao J Y, Yang X S and Xu C S. 2021.Multimodal-based zero-shot human action recognition. Journal of Image and Graphics,26(07):1658-1667(吕露露,黄毅,高君宇,杨小汕,徐常胜. 2021. 多模态零样本人体动作识别. 中国图象图形学报,26(07):1658-1667)[DOI:10.11834/jig.200503]

Ma Yuanbo, Xu Xing, Shen Fumin and Shen Heng Tao.2020.Similarity preserving feature generating networks for zero-shot learning.Journal of the 2020 Neurocomputing, 406: 333-342[DOI: https://doi.org/10.1016/j.neucom.2019.08.111]

Mishra Ashish, Reddy Shiva Krishna, Mittal Anurag and Murthy Hema A.2018.A Generative Model for Zero Shot Learning Using Conditional Variational Autoencoders//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.Salt Lake City, UT, USA: IEEE:2269-22698[DOI: 10.1109/CVPRW.2018.00294]

Nilsback Maria-Elena and Andrew Zisserman.2008.Automated Flower Classification over a Large Number of Classes//Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing.Bhubaneswar, India: IEEE: 722-729[DOI: 10.1109/ICVGIP.2008.47]

Patterson G, J.and Hays.2012.SUN attribute database: Discovering, annotating, and recognizing scene attributes//Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition.Providence, RI, USA: IEEE: 2751-2758[DOI: 10.1109/CVPR.2012.6247998]

Rohit Keshari, Richa Singh and Mayank Vatsa.2020.Generalized Zero-Shot Learning via Over-Complete Distribution//Proceedings of the 2020 IEEE/CVF International Conference on Computer Vision and Pattern Recognition.Seattle, WA, USA: IEEE: 13297-13305[DOI: 10.1109/CVPR42600.2020.01331]

Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees G. M.Snoek and Ling Shao.2020.Latent Embedding Feedback and Discriminative Features for Zero-Shot Classification//Vedaldi, A., Bischof, H., Brox, T., Frahm, JM.eds.European Conference on Computer Vision.Glasgow:Springer: 475-495[DOI: 10.1007/978-3-030-58542-6_29]

Schonfeld E, Ebrahimi S, Sinha S.2019.Generalized zero-and few-shot learning via aligned variational autoencoders//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Long Beach, CA, USA:IEEE:8239-8247[DOI: 10.1109/CVPR.2019.00844]

Simonyan, Karen and Zisserman and Andrew.2014.Very deep convolutional networks for large-scale image recognition [DB/OL].[2014-09-04].https://arxiv.org/pdf/1409.1556.pdf

Sohn, Kihyuk and Yan, Xinchen and Lee, Honglak.2015.Learning Structured Output Representation Using Deep Conditional Generative Models.Advances in neural information processing systems, 3483-3491[DOI: 10.5555/2969442.2969628]

Sung Flood, Yang Yongxin, Zhang Li, Xiang Tao, Philip H.S.Torr, Timothy M and Hospedales.2018.Learning to Compare: Relation Network for Few-Shot Learning//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Salt Lake City, UT, USA: IEEE: 1199-1208[DOI: 10.1109/CVPR.2018.00131]

Tong Bin, Wang Chao, Martin Klinkigt, Yoshiyuki Kobayashi and Yuuichi Nonaka.2019.Hierarchical Disentanglement of Discriminative Latent Features for Zero-Shot Learning//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Long Beach, CA, USA: IEEE: 11459-11468[DOI: 10.1109/CVPR.2019.01173]

Verma Vinay Kumar, Arora Gundeep, Mishra Ashish and Rai Piyush.2018.Generalized Zero-Shot Learning via Synthesized Examples//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Salt Lake City, UT, USA: IEEE: 4281-4289[DOI: 10.1109/CVPR.2018.00450]

Wah C. , Branson S. , Welinder P. , Perona P. and Belongie S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. CNS-TR-2011-001. California Institute of Technology

Wang Wenlin, Xu Hongteng, Wang Guoyin, Wang Wenqi and L-awrence Carin. 2021.Zero-Shot Recognition via Optimal Tra-nsport//Proceedings of the 2021 IEEE Winter Conference onApplications of Computer Vision.Waikoloa, HI, USA: IEEE:3470-3480[DOI: 10.1109/WACV48630.2021.00351]

Wang Yaqing and Yao Quanming and Kwok, James T.and Ni, Lionel M.2020.Generalizing from a Few Examples: A Survey on Few-shot Learning.ACM Computing Surveys, 53:1-34[DOI: 10.1145/3386252]

Xian Yongqin, Saurabh Sharma, Bernt Schiele and Zeynep Akata.2019.F-VAEGAN-D2: A Feature Generating Framework for Any-Shot Learning//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Long Beach, CA, USA: IEEE: 10267-10276[DOI: 10.1109/CVPR.2019.01052]

Xian Yongqin, Tobias Lorenz, Bernt Schiele and Zeynep Akata. 2018.Feature Generating Networks for Zero-Shot Learning//Proceedings of the 2018 IEEE/CVF International Conference on Computer Vision and Pattern Recognition.Salt Lake City, UT, USA: IEEE: 5542-5551[DOI: 10.1109/CVPR.2018.00581]

Xian Yongqin, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein and Bernt Schiele.2016.Latent Embeddings for Zero-Shot Classification//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition.Las Vegas, NV, USA: IEEE: 69-77[DOI: 10.1109/CVPR.2016.15]

Ye Zihan, Lyu Fan, Li Linyan, Fu Qiming, Ren Jinchang, Hu Fuyuan.2019.SR-GAN: SemanticRectifying Generative Adversarial Network for Zero-shot Learning//Proceedings of the 2019 IEEE International Conference on Multimedia and Expo.Shanghai, China:IEEE:85-90[DOI: 10.1109/ICME.2019.00023]

Zhao Peng, Wang Chunyan, Zhang Siying and Liu Zhengyi. 2021. A Zero-Shot Image Classification Method Based on Subspace Learning with the Fusion of Reconstruction. Journal of Computers, 44(02):409-421(赵鹏，旺纯燕，张思颖，刘政怡. 2021. 一种基于融合重构的子空间学习的零样本图像分类方法. 计算机学报，44(02):409-421)[DOI:10.11897/SP.J.1016.2021.00409]

Zhu Yizhe, Mohamed Elhoseiny, Liu Bingchen, Peng Xi and Ahmed Elgammal.2018.A Generative Adversarial Approach for Zero-Shot Learning from Noisy Texts//Proceedings of the 2018 IEEE/CVF International Conference on Computer Vision and Pattern Recognition.Salt Lake City, UT, USA: IEEE:1004-1013[DOI: 10.1109/CVPR.2018.00111]

作者简介

韩阿友，1997 年生，男，硕士研究生，主要研究方向为机器学习和计算机视觉。
E-mail: you1120@zut.edu.cn

杨关，通信作者，1974 年生，男，副教授，博士，

主要研究方向为机器学习和计算机视觉。
E-mail: yangguan@zut.edu.cn
刘小明，男，讲师，博士，主要研究方向为自然语言处理。
E-mail: ming616@zut.edu.cn
刘阳，男，讲师，博士，主要研究方向为机器学习和模式识别。
E-mail: yangl@xidian.edu.cn