计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.181202316

融合三元卷积神经网络与关系网络的小样本食品图像识别

吕永强[1,2]　闵巍庆[2]　段　华[1]　蒋树强[2]
1 山东科技大学数学与系统科学学院　山东 青岛 266590
2 中国科学院计算技术研究所　北京 100190

摘　要　食品识别在食品健康和智能家居等领域获得了广泛关注。目前大部分的食品识别工作是基于大规模标记样本的深度神经网络,这些工作无法有效地识别只有少量样本的类别,因此小样本食品识别是一个亟待解决的问题。目前基于度量学习的小样本识别方法着重于探究样本之间的相似度信息,忽略了类内与类间更加细粒度的区分。学习类内与类间区分信息的主流方法是基于线性度量函数的三元卷积神经网络,然而对于食品图像而言,线性度量函数的鉴别能力不足。为此,引入可学习的关系网络作为三元卷积神经网络的非线性度量函数,进一步提出了一种基于非线性度量的三元神经网络用于小样本食品识别方法。该方法使用三元神经网络学习图像的特征嵌入表示,然后采用鉴别能力更强的关系网络作为非线性度量函数,基于端到端的训练方式来学习类内与类间更加细粒度的区分信息。此外,提出了一种可以使模型训练更加稳定的三元组样本在线采样方案。通过在 Food-101,VIREO Food-172 和 ChineseFoodNet 食品数据集上的实验结果可知,相比基于孪生网络的小样本学习方法,所提方法的性能平均提高了3.0%,相比基于线性度量函数的三元神经网络的方法,所提方法的性能平均提升了 1.0%。文中还探究了损失函数的阈值、三元组采样的参数和初始化方式对实验性能的影响。

关键词:食品识别;小样本识别;非线性度量;三元神经网络
中图法分类号　TP391

Few-shot Food Recognition Combining Triplet Convolutional Neural Network with Relation Network

LV Yong-qiang[1,2],MIN Wei-qing[2],DUAN Hua[1] and JIANG Shu-qiang[2]
1 College of Mathematics and System Science,Shandong University of Science and Technology,Qingdao,Shandong 266590,China
2 Institute of Computing Technology,Chinese Academy of Sciences,Beijing 100190,China

Abstract　Food recognition attracts wide attention in the fields of food health and smart home. Most existing work focuses on food recognition with large-scale labeled samples,thus failing to robustly recognize food categories with few samples,under this condition,few-shot food recognition is an urgent problem. Most metric learning based few-shot recognition methods emphasize more on the similarity values of the image pairs without paying substantial attention to the inter-class and intra-class variations. Most works mainly use triplet convolutional neural network with linear metric function to learn the inter-class and intra-class information,however the liner metric function is not discriminative enough for measuring similarities of food images. To address this problem,this paper used the learnable relation network as non-linear metric and proposed a triplet network with relation network to solve the above two disadvantages of the few-shot learning and triplet network. This model adopts triplet network as feature embedding network for the image feature learning and uses a relation network with better discrimination as the non-linearity metric to learn the inter-class and intra-class information. Also the proposed model is trained end-to-end. In addition,this paper proposed an on-line mining rule for triplet samples,which makes the model stable in the training stage. The comprehensive experimental was conducted on three food datasets,which are Food-101,VIREO Food-172 and ChineseFoodNet. Compared with popular few-shot learning methods,such as Relation network,Matching network,the proposed model achieves an average improvement of about 3.0%,and compared with triplet network with liner metric,it achieves an average improvement of about 1.0%. Also this paper explored the influence of the margin in the loss function,parameters setting of online triplet sampling and initialization

到稿日期:2018-12-14　返修日期:2019-04-29　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61532018,61602437);教育部人文社会科学研究项目(18YJAZH017);山东省自然科学基金(ZR2017MF027);山东科技大学领军人才与优秀科研团队计划资助项目(2015TDJH102)
This work was supported by the National Natural Science Foundation of China (61472229,61602279,71704096,31671588),Sci. & Tech. Development Fund of Shandong Province of China (2016ZDJS02A11,ZR2017BF015,ZR2017MF027),Humanities and Social Science Research Project of the Ministry of Education (16YJCZH154,16YJCZH041 16YJCZH012,18YJAZH017),Taishan Scholar Climbing Program of Shandong Province,and SDUST Research Fund (2015TDJH102).
通信作者:段华(huaduan59@163.com)

methods on experiment performance.

Keywords Food recognition, Few-shot learning, Non-linear metric, Triplet network

1 引言

食品识别是计算机视觉[1]、数据挖掘[2]以及多媒体社交[3]等领域的重要研究课题，在食品自动化检测[4]、食品管理、食品安全、食品的趋势和流行性分析以及智能家居中有着广泛的应用，例如智能厨房和智能个人营养日志[5]等。

食品识别面临许多挑战。据维基百科统计，食品类别已经超过8000种[6]，从现实世界收集的食品数据集符合典型的长尾分布，即许多不常见的食品类别只能收集到少量的样本。一个鲁棒的食品识别模型，不仅可以识别常见的食品类别，还可以有效地识别不常见的食品类别。现有基于深度神经网络的食品识别方法[6-9]需要大规模的标注样本才能进行有效的模型训练，对于只有少量标注样本的食品类别，这些方法的识别能力很差。因此，本文主要探究面向小样本的食品识别问题。

近年来，学者们对小样本学习的兴趣越来越浓厚[10-15]，小样本学习主要应用于字符识别[10]、图像识别[11-12]和图像分割[15]等领域。小样本学习的目标是学习识别一个只有少量标记样本的类别。目前基于小样本学习的工作是利用图像对之间的相似度对图像进行分类，即相似度大的为同类，相似度小的为异类。这些方法忽视了图像对在类内与类间更加细粒度的对比信息。如图1所示，图1(a)是“糖醋鱼”类别的两个样本，图1(b)是“水煮鱼”和“水煮肉”同类图像对。对于这类视觉上很相似的异类图像对，这些模型由于忽略了更加细粒度的区分信息，因此识别能力较弱。

食品类别：糖醋鱼　　食品类别：糖醋鱼

(a)

食品类别：水煮肉　　食品类别：水煮肉

(b)

图1 食品图像对比

Fig. 1 Comparison of food images

与其他类型数据集相比，食品识别面临更多挑战：1)食品图像的识别属于细粒度分类[7]，食品图像在类内和类间的细粒度的区分对食品识别十分重要；2)许多食品图像没有固定的空间布局，即没有固定的结构，如“糖醋鱼”“沙拉”等。对于小样本食品识别而言，由于训练集与测试集的类别相关性较弱，利用训练集获得更具有类别区分度的信息对于提升模型对新类别的识别能力显得更加困难。为此，我们将三元卷积神经网络用于小样本食品的识别，以学习更加细粒度的区分信息。

三元卷积神经网络[16]主要用于人体重识别[17-18]、车辆重识别[19]等领域。相比孪生网络(连体网络)，三元卷积神经网络通过同时控制类内和类间图像对的相似度差异，来学习更加具有类别区分度的特征表示。然而大多数基于三元卷积神经网络的工作[16-19]是使用固定的线性度量，例如余弦距离或者欧氏距离等，这种利用固定距离作为度量函数的方法存在以下3个缺陷：1)模型过度依赖特征学习网络的学习能力，即整个模型在学习过程中，会受到特征学习网络的区分能力的限制；2)对于需要细粒度区分的食品图像而言，线性的度量方法的辨别能力不够强；3)固定的度量算法需要根据不同的网络模型以及不同的数据集进行人工选择，无法自适应地根据数据集和网络结构进行学习。最近，Sung等[12]提出了一种基于卷积神经网络的关系学习网络，此网络可以根据不同的数据集和模型自适应地学习一个非线性度量函数，通过损失函数可以同时调节特征嵌入网络和关系网络，使模型能够取得更好的性能。

此外，三元卷积神经网络的训练条件十分苛刻。在样本空间中构建的三元组中存在许多对于训练不利的三元组，如过于容易训练的三元组和过于难训练的三元组。过于容易训练的三元组会使损失函数接近0甚至等于0，导致网络学习过慢或者不学习；过于难训练的三元组会使损失函数过大，导致模型收敛慢甚至不收敛。为此，Hermans等[17]提出了一种在线的三元组采样方案，称为“batch hard”，该方法旨在从一个批量的样本中选择出最难训练的三元组。然而基于非线性度量的三元卷积神经网络在训练过程中，即使采用“batch hard”的三元组筛选方案，仍然存在大量对模型训练不利的三元组。

针对上述不足，我们提出了一个基于非线性度量学习的三元神经网络用于小样本食品识别。该模型主要由两部分组成：用于图像特征学习的特征嵌入子网络和用于非线性度量函数学习的关系学习子网络，如图2所示。其中特征嵌入子网络采用3个参数共享的VGG16，分别提取最后一个卷积层的特征作为三元组的特征嵌入表示；利用两个参数共享的多层神经网络构建可学习的非线性度量函数，用于正负样本图像对的关系学习；最终采用端到端的训练方式学习整个模型。此外，为了模型训练的稳定性，本文提出了一种新的三元组在线采样方案，称为“limited batch hard”，其利用正负样本图

像对的关系得分，挑选出适合模型训练的三元组。在 Food-101[1]，VIREO Food-172[20] 和 ChineseFoodNet[21] 这 3 个食品数据集上，相较现有方法，本文所提模型取得了最高性能。

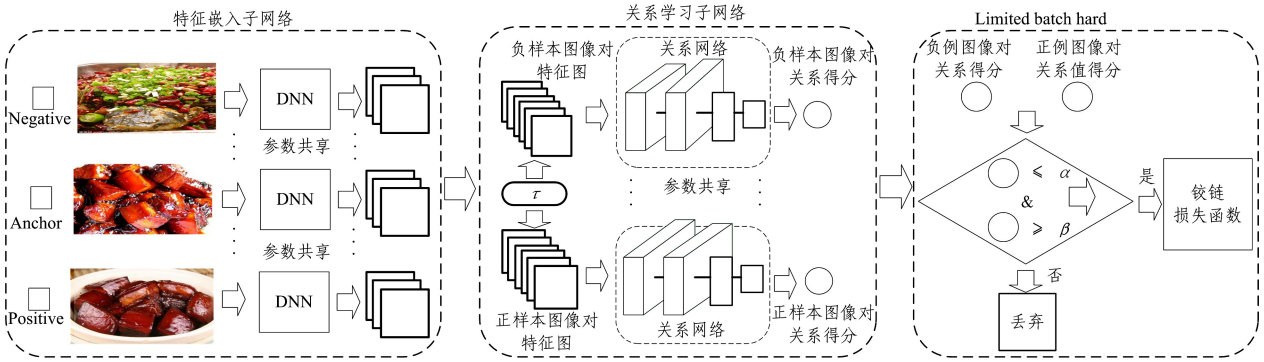

图 2 基于非线性度量学习的三元卷积神经网络框架图

Fig. 2 Architecture of triplet convolutional neural network with nonlinear metric network

2 相关工作

2.1 食品识别

由于食品图像有较大的外观差异，这种差异不仅来自光照和焦点的变化，还来自不同烹饪方法、切割方式或者不同的食材搭配造成的形状、纹理和其他视觉属性的变化，因此食品识别成为计算机识别领域的难题之一。Min 等[22] 针对不同的面向食品的任务和应用进行了系统的总结。与传统的人工设计的特征相比，Kawano 等[23] 发现深度特征明显优于人工设计的特征；Kagaya 等[24] 进一步提取深度的视觉特征用于食品的检测和识别。上述工作只考虑了食品图像的视觉信息。许多工作关注面向餐馆上下文信息的食品识别，如 Xu 等[25] 探索餐馆的地理位置信息和菜单信息简化食品的分类问题。此外，一些工作关注使用食品的原料信息和食谱信息构建多模态的识别或检索模型，以提高识别和检索性能。如 Min 等[26] 提出了一个多模态多任务模型，该模型同时学习食品图像的视觉特征表示和食品的原料信息的特征表示，并利用这两种信息提升模型的识别能力。Min 等[27] 利用丰富的食谱信息和食品图像的视觉信息，提出了一个多模态食品检索模型。一些工作[28] 进一步考虑了食品的原料信息，利用多任务的方法建模原料信息与视觉信息之间的关联，提出了一个多任务识别模型。此外，一些工作关注到图像背景噪音的干扰，Mei 等[29] 使用 Faster R-CNN 提取图像的判别性区域特征，避免噪音的干扰且使提取的特征更具判别性。上述工作都是在大规模数据的基础上建立模型，而本文研究面向小样本的食品识别问题。

2.2 小样本识别

近年来，学者们采用不同的策略处理小样本问题，主要有基于度量学习的小样本学习方法[10-12] 和基于元学习[13-14] 的小样本学习方法。由于基于度量学习的小样本学习方法能取得较好的性能，因此本文主要关注该方法。Koch 等[10] 提出使用孪生网络作为图像特征嵌入网络，通过学习特征嵌入网络获得图像对的特征表示，使其能够通过固定的度量算法实现图像的分类。在孪生网络的基础上，Vinyals 等[11] 提出 Matching network，通过构建支持集和查询集，并将其嵌入到一个共享的特征空间中，利用余弦距离计算损失函数来学习整个网络。Siamese network 和 Matching network 都采用了线性度量方法，Sung 等[12] 提出了一种基于非线性度量方法的 Relation network，利用可学习的卷积神经网络作为非线性度量函数。相比线性度量方法，Relation network 取得了更好的性能。然而，上述工作没有考虑图像对在类内与类间信息之间的差异，对于相似度小的同类图像对和相似度大的异类图像对，这些模型的识别能力较弱。我们采用三元卷积神经网络代替孪生网络以学习图像对在类内与类间更加细粒度的区分信息，同时融合非线性关系网络，提出了一个端到端训练的小样本学习模型。

3 小样本食品识别模型

3.1 问题定义

小样本学习问题中，训练与测试样本通常由一系列训练集与测试集构成。假设有 C 个训练类别，共有 N 个有标注的训练样本，定义训练集 $D_{base}=\{(x_i^{base},y_i^{base})\}_{i=1}^{N}$，$y_i^{base}\in\{B_1,B_2,\cdots,B_C\}$，其中 x_i^{base} 是指采样的图像，y_i^{base} 是指 x_i^{base} 的标签。对于测试集，假设有 L 个新的类别和 M 个测试样本，定义测试样本集 $D_{novel}=\{(x_i^{novel},y_i^{novel})\}_{i=1}^{M}$，其标签集为 $y_i^{novel}\in\{N_1,N_2,\cdots,N_L\}$。值得注意的是，训练集与测试集的样本空间是完全不相关的。

对于小样本学习，首先定义支持集和查询集。以训练集为例，随机从 D_{base} 采样 C 个类别，并从每个类别中随机采样 K 个样本构成支持集 $S=\{(x_i^{base},y_i^{base})\}_{i=1}^{m}$，$m=(C\times K)$。定义查询集 $S=\{(x_i^{base},y_i^{base})\}_{i=1}^{n}$，从支持集的 C 个类别中随机选择一个类别，并从选中的类别中随机采样 n 个样本。如果支持集中包含 C 个不同的类别，且每个类别包含 K 个样本，则称该任务为“C-way K-shot”。一般而言，在小样本学习的设置中 K 往往是很小的，例如 $K=1$ 或者 $K=5$。基于“C-

way K-shot"任务的目标是提供一张查询图像$\hat{x}$，利用支持集学习一个分类映射 $c_S(\hat{x})$，从而得到查询类别的概率分布 $P(\hat{y}|\hat{x},S)$，其中 $\hat{y}$ 为预测的标签。

3.2 模型结构

基于非线性度量学习的三元神经网络主要由两个子网络构成：图像嵌入网络和关系学习网络。

3.2.1 图像特征嵌入网络

基于三元卷积神经网络的图像特征嵌入网络 f_θ 是受到孪生网络（连体网络）的启发，其结构由 3 个参数共享的前馈深度神经网络构成，本文采用 3 个 VGG16 深度神经网络作为特征嵌入网络，分别用于三元组 $X=\{x^-,x,x^+\}$ 的特征提取。具体而言，提供 3 个图像输入 x^-,x,x^+，分别表示 negative 图像、anchor 图像和 positive 图像，其中 x 与 x^+ 属于同一类别的样本，x^- 与 x 属于不同类别的样本。已有工作[16-18]采用分类层之前的全连接层作为图像的嵌入表示，然后利用固定距离算法（如 L_2 距离）得到三元组的距离表示：

$$Triplet(x^-,x,x^+)=\begin{bmatrix}\|(f_\theta(x),f_\theta(x^-))\|_2\\ \|(f_\theta(x),f_\theta(x^+))\|_2\end{bmatrix} \tag{1}$$

其中，f_θ 为样本的特征嵌入表示，θ 为特征嵌入网络的参数。本文特征嵌入表示为卷积层提取的特征图，其可以契合关系学习网络的输入，而且相较全连接层，卷积层的特征包含更加丰富的图像信息。

3.2.2 关系学习网络

如图 3 所示，关系学习网络 g_φ 包含 2 个卷积块和 2 个全连接层。每个卷积块由 64 个 3×3 的卷积核构成的卷积层和 1 个 2×2 的最大池化层组成，在卷积层中使用批正则化处理和 ReLU 非线性激活。两个全连接层的特征输出的维度分别为 8 维和 1 维。第一个全连接层使用 ReLU 函数作为激活函数，第二个全连接层使用 Sigmoid 函数作为激活函数。

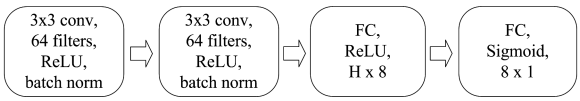

图 3 关系网络结构

Fig. 3 Structure of relation network

为契合关系网络的输入，我们将一个算子 τ 用于特征图的融合，本文采用的是特征图的深度拼接，例如从最后一个卷积层提取的特征图维度为 $14\times14\times512$，通过正样本或者负样本图像对进行特征融合，融合特征图维度为 $14\times14\times1024$，则最终的三元组的特征表示为：

$$Triplet(x^-,x,x^+)=\begin{bmatrix}(\tau(f_\theta(x),f_\theta(x^-))\\ \tau(f_\theta(x),f_\theta(x^+))\end{bmatrix} \tag{2}$$

将深度融合后的正负样本图像对的特征图分别输入到两个参数共享的关系学习网络，得到两个范围在 0 到 1 的关系值来表示正负样本图像对之间的相似度，称其为非线性关系得分。本文认为非线性关系得分越高，两张图像就越相似，相反关系值越低，两张图像就越不相似。最终的关系得分表示为：

$$r(x^-,x,x^+)=\begin{bmatrix}g_\varphi(\tau(f_\theta(x),f_\theta(x^-)))\\ g_\varphi(\tau(f_\theta(x),f_\theta(x^+)))\end{bmatrix} \tag{3}$$

其中，r 为关系值得分，$g_\varphi(\cdot)$ 为关系网络，φ 为关系网络的参数，τ 为用于融合特征的算子，$\tau(f_\theta(x),f_\theta(x^-))$ 为负样本图像对的特征深度融合，$\tau(f_\theta(x),f_\theta(x^+))$ 为正样本图像对的特征深度融合。

3.3 端到端的联合优化

与先前的工作不同，本文没有选择用于学习两张图像相似度的损失函数，例如交叉熵损失函数或者均方误差损失函数等，而是使用能够学习三张图像区分信息的铰链损失函数。铰链损失函数通过同时限制正负样本关系值可以使模型学习得到更加具有区分度的信息。具体而言，当输入一个三元组 (x^-,x,x^+) 时，如前文中提到的，x 与 x^+ 属于同类，x 与 x^- 属于异类，将三元组输入的图像特征嵌入网络得到特征图分别为 $f_\theta(x^-),f_\theta(x),f_\theta(x^+)$，经过算子 τ 得到融合后的正样本图像对和负样本图像对的特征图分别为 $\tau(f_\theta(x),f_\theta(x^+))$ 和 $\tau(f_\theta(x),f_\theta(x^-))$。在训练阶段将融合的特征图 $\tau(f_\theta(x),f_\theta(x^-))$ 和 $\tau(f_\theta(x),f_\theta(x^+))$ 分别输入到关系学习网络 g_φ，最终的铰链损失函数定义为：

$$L_{triplet}(x,x^-,x^+)=\max\{m+g_\varphi(\tau(f_\theta(x),f_\theta(x^-)))-g_\varphi(\tau(f_\theta(x),f_\theta(x^+)))\} \tag{4}$$

其中，m 为阈值，铰链损失函数使得正样本图像对之间的关系值得分大于负样本图像对之间的关系值得分。铰链损失函数不仅指导特征嵌入模型产生图像的嵌入，还指导关系网络学习。

3.4 "limited batch hard"三元组挑选方案

三元卷积神经网络的模型训练条件十分苛刻，首先随着数据集的增大，三元组的采样空间远大于样本空间，训练的时间将变得冗长。其次在构建的三元组中有很多对训练不利的三元组，如容易训练的三元组，即正样本图像对的关系值很大、负样本图像对的关系值很小、铰链损失函数的值接近 0 或者等于 0，这些三元组对模型的调节能力很小或者不调节模型。更糟糕的是，如果选取过于难训练的三元组，网络将很难收敛，以至于无法学习到具有区分性的信息。因此，三元组的选择直接关系到三元卷积神经网络训练的稳定性。Hermans[17]等提出了一种三元组挑选方案，称为"batch hard"。具体而言，对于每个批量数据，随机选择 P 个类别，然后从每个类别中随机选择 K 个样本，这样每个批量样本中就含有 PK 个图像。对于每个样本 x 在构建三元组时可以选择出最难训练的正样本图像对和最难训练的负样本图像对，最终筛选出 PK 个三元组。因此，基于"batch hard"的关系网络的损失函数为：

$$L_{BH}((\theta,\varphi);x)=\sum_{i=1}^{P}\sum_{a=1}^{K}[\{m-\max_{p=1,\cdots,K}g_\varphi(\tau(f_\theta(x_a^i),f_\theta(x_p^i)))+\min_{\substack{j=1,\cdots,P\\ n=1,\cdots,K\\ j\neq i}}g_\varphi(\tau(f_\theta(x_a^i),f_\theta(x_n^i)))\}] \tag{5}$$

但是利用“batch hard”的方法筛选基于关系网络的三元组，仍然存在很多对训练不利的三元组。具体而言，关系网络最后一层全连接的输出经过 Sigmoid 函数后被规范到 0 到 1 之间，我们认为输出值越接近 1，图像对就越相似，越接近 0，图像对就越不相似。我们发现在一个批量的样本中存在很多不利于训练的三元组，例如有许多正样本图像对的关系值在 0.01 左右，且负样本的关系值在 0.9 左右。如果直接用这些三元组训练网络，会造成网络训练不稳定或者不收敛。

如图 4 所示，本文提出一个新的三元组采样方案。与“batch hard”方法相似，对于一个样本 x，从其构成的三元组中采样，但是在采样最难训练的三元组之前，“limited batch hard”需要对所有的正样本图像对和负样本图像对进行限制，例如只选取正样本图像对关系得分大于等于 α，同时负样本图像对的关系得分小于 β 的三元组。基于新的三元组采样规则，可以采样出对训练有益的三元组使模型可以稳定地训练。最终，基于新的三元组采样方案的损失函数为：

$$L_{BH}((\theta,\varphi);x)=\sum_{i=1}^{P}\sum_{a=1}^{K}[\{m-\max_{p=1,\cdots,K}g_{\varphi}(\tau(f_{\theta}(x_a^i),f_{\theta}(x_p^i)))+\min_{\substack{j=1,\cdots,P\\n=1,\cdots,K\\j\neq i}}g_{\varphi}(\tau(f_{\theta}(x_a^i),f_{\theta}(x_n^j)))\}]$$

$$\text{s.t. } g_{\varphi}(\tau(f_{\theta}(x_a^i),f_{\theta}(x_n^j)))\geqslant\alpha$$

$$g_{\varphi}(\tau(f_{\theta}(x_a^i),f_{\theta}(x_p^i)))\leqslant\beta \tag{6}$$

其中，α 为负样本关系值的阈值，β 为正样本关系值的阈值。

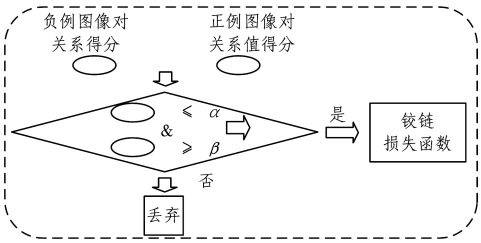

图 4 limited batch hard 的流程图

Fig. 4 Flow diagram of limit batch hard

3.5 模型的算法流程

本节将详细介绍模型的算法流程。

Step1 模型初始化与三元组生成

采用在训练集上预训练的特征提取模型和关系学习模型分别初始化特征嵌入网络和关系学习网络。

从训练集中随机选取 P 个类别，并从每个类别随机采样 K 个样本。将每个样本作为 anchor，分别构建三元组作为输入，最终获得 PK 个三元组。

Step2 获得三元组特征嵌入表示

Step2.1 获取样本特征表示

将每个三元组分别输入到 3 个参数共享的特征嵌入网络，分别提取最后一个卷积层的特征图作为最后的特征嵌入表示 $f_\theta(x^-)$，$f_\theta(x)$，$f_\theta(x^+)$。

Step2.2 融合正负样本的特征

利用算子 τ 分别融合正负样本图像对的特征，得到特征嵌入表示为 $\tau(f_\theta(x),f_\theta(x^+))$，$\tau(f_\theta(x),f_\theta(x^-))$。其中，算子 τ 采用的是特征的深度融合。

Step3 计算正负样本图像对的关系值得分

将融合后的正负样本的特征嵌入表示分别输入到关系学习网络 g_φ，分别得到关系值得分 $g_\varphi(\tau(f_\theta(x),f_\theta(x^+)))$ 和 $g_\varphi(\tau(f_\theta(x),f_\theta(x^-)))$。

Step4 三元组筛选和模型更新

利用“limited batch hard”三元组筛选规则，首先确定阈值 α 与 β，如果 $g_\varphi(\tau(f_\theta(x),f_\theta(x^+)))\leqslant\beta$ 且 $g_\varphi(\tau(f_\theta(x),f_\theta(x^-)))\geqslant\alpha$，则利用三元组的关系值得分计算损失函数，并更新模型，否则丢弃该三元组。

4 实验及结果分析

本节首先介绍实验数据集以及实验的实现细节，然后验证本文所提模型在不同数据集上的实验性能，最后分析模型初始化对性能的影响、铰链函数阈值和“limited batch hard”参数的敏感度。

4.1 数据集及划分

为验证方法的有效性，本文在 3 个重要的食品数据集：Food-101[1]，VIREO Food-172[20] 和 ChineseFoodNet[21] 上进行了全面的对比实验。

3 个数据集的部分样本如图 5 所示。

图 5 3 个数据集的样例

Fig. 5 Samples of three food datasets

1)Food-101：数据集包含 101 类食品类别，每个类别包含 1000 张食品图像，共有 101000 张食品图像，其中大多数食品为西方菜品。由于小样本学习的训练集与测试集是完全不相交的，因此本文随机分割数据集，其中 71 类为训练集，30 类为测试集。

2)VIREO Food-172：数据集包含 172 个食品类别，共有 110241 张食品图像。所有的食品图像都是中国菜，并且所有的食品图像都是从百度和谷歌搜索中获取的。同样，本文随机分割数据集，其中 132 类作为训练集，40 类作为测试集。

3)ChineseFoodNet：数据集主要由许多不同烹饪风格的中国菜品构成，包含 208 个食品类别，共有 185 628 张食品图像。本文随机分割数据集，其中 158 类作为训练集，50 类作为测试集。

4.2 实现细节

本节主要介绍模型的初始化方式以及模型的参数设置。

4.2.1 模型的初始化

在实验过程中，模型初始化对于模型训练至关重要。首先模型初始化对三元组采样有着决定性的作用，关系网络作为采样三元组的度量方法，如果使用随机初始化，则无法挑选出适合训练的三元组。同样，图像的嵌入网络作为关系网络提供输入的关键，图像嵌入网络的初始化也需要符合三元组初始筛选的条件。为此，我们分别对图像嵌入网络和关系网络采用合适的预训练模型初始化。

对于图像嵌入网络，初始化是采用基于训练集得到的分类模型参数，目的是获得具有区分性的特征。具体方法如下：首先使用基于 ImageNet 数据集预训练的模型[30]初始化 VGG16 深度神经网络，然后利用训练集数据训练一个分类模型，最后得到一个具有训练集类别区分度的分类模型。

对于关系网络的初始化，首先利用经过训练集微调过的 VGG16 分类模型，分别提取出训练集与测试集的最后一个卷积层的特征，在基于“5-way 1-shot”的训练机制下使用均方误差损失训练关系网络，最终利用测试集上准确率最高的一组参数作为非线性关系学习网络的初始化。采用此初始化方案的优点如下：1）由于三元组的筛选是由特征嵌入网络的特征与关系网络的判别性能决定的，因此获得更具判别性的特征和具有类别区分度的非线性分类器，有利于基于非线性规则的三元组的在线采样；2）在整个预训练过程中，本文只使用了训练集的样本，测试集信息没有流入整个模型中，因此模型的初始化参数符合小样本学习对数据集的设置要求；3）对于随机初始化模型，整个训练模型的过程需要大量时间。由于使用预训练的参数初始化模型，整个模型已经具有较好的类别区分能力，因此在训练过程中仅使用较小学习率微调网络模型就可以很快地达到更好的收敛效果。

4.2.2 参数设置

对于模型的优化算法，我们采用的是 Adam[31]优化算法，相比随机梯度下降，Adam 优化算法的收敛速度更快且更稳定。实验中使用 Adam 优化算法的默认超参数设置（$\beta_1=0.9$，$\beta_2=0.999$，$\varepsilon=10^{-8}$）。

本文只考虑“5-way 1-shot”的训练机制。由于网络模型的复杂度较高，将批样本大小设置为 49，包含 $P(P=7)$ 个类别以及每个类别随机选取的 $K(K=7)$ 个样本。由于模型初始参数是采用预训练的参数，因此训练过程使用较低的初始学习率，即 10^{-5}。训练过程中微调图像特征嵌入网络最后一个卷积层以及整个非线性关系学习网络。对于其他实验参数设置，铰链损失函数的阈值为 0.9，“limited batch hard”参数设置为(0.4，0.6)。为防止过拟合现象，在每个卷积层中加入批正则化处理和 ReLU 非线性层，全连接层的 dropout 率设置为 0.5，模型在训练迭代 20000 次左右就可以收敛。

4.3 实验结果与分析

4.3.1 方法对比实验

为验证本文模型的性能，分别在 Food-101，VIREO Food-172 与 ChineseFoodNet 3 个数据集上，选取流行的小样本方法 Siamese network[10]、Matching network[11]、Relation network[12]和基于平方欧氏距离的三元卷积神经网络[32]作为基准方法，进行全面的对比实验。由于文献[32]面向的是手写字体识别，其构建的网络不适合食品图像识别，因此使用 VGG16 代替原文中的网络结构。

本文算法与其他算法的对比如表 1 所列。由表 1 可以得到：1）基于三元卷积神经网络的方法比基于孪生网络的小样本方法可以取得更好的性能。其主要原因是三元卷积神经网络相比孪生网络可以学习到更加细粒度的类别差异信息，可以提取更加具有辨识度的特征，因此模型的识别性能更强。具体而言，对比 3 个数据集的实验结果可知，所提方法比 Siamese network 的性能分别提高 7.8%，16.6%和 15.6%，比 Matching network 的性能分别提高 11.3%，1.6%和 17.2%，比 Relation network 的性能分别提高 3.0%，2.9%和 2.3%。2）关系网络作为可学习的非线性度量函数，可以根据数据集和特征嵌入的网络结构自适应学习，其识别能力更强。与基于平方欧氏距离的三元卷积神经网络相比，所提模型性能分别提升 0.8%，1.7%和 3.7%。所有实验结果表明，相比基于孪生网络的方法和基于线性距离的三元卷积神经网络，基于非线性度量学习的三元神经网络可以取得更好的性能。

表 1 不同小样本方法在不同数据集上的实验结果

Table 1 Experimental results of different datasets

（单位：%）

模型	度量方法	Food-101	VIREO Food-172	ChineseFoodNet
Siamese network[10]	余弦距离	49.1	60.3	50.5
Matching network[11]	余弦距离	45.6	73.6	48.9
Relation network[12]	非线性距离	53.9	74.0	63.8
Triplet network[32]	平方欧氏距离	56.1	75.2	62.4
Our Method	非线性距离	56.9	76.9	66.1

4.3.2 铰链函数阈值 m 的敏感度分析

为了探究阈值对实验的影响，我们设计基于不同阈值的对比实验。如图 6 所示，随着阈值的逐渐提升，模型的识别能力呈现出先提升后下降的趋势，当阈值为 0.9 时，模型的性能达到最高。阈值作为铰链损失函数的一部分，控制着正样本图像对与负样本图像对的收敛性。如果阈值设置得过小，则正样本图像对与负样本图像对之间的区分度不够，很难通过训练得到具有区分性的模型；如果阈值设置得过大，则模型很难收敛。

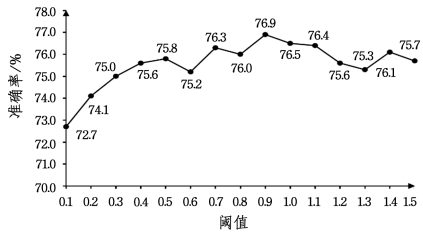

图 6 不同阈值的实验结果

Fig. 6 Experimental results of different margins

4.3.3 三元组采样参数 α 和 β 的敏感度分析

本节在 VIREO Food-172 数据集上分析三元组采样方案

的参数敏感度，通过设置“limited batch hard”中阈值 α 和 β 的取值，不断地缩小三元组采样的范围。本文通过对比“batch hard”的采样方案来验证提出的采样规则的有效性，实验结果如表 2 所列。

表 2 三元组采样的对比实验结果

Table 2 Experimental results of triplet sampling

（单位：%）

模型	准确率
“batch hard”[17]	56.2
(0.1,0.9)	72.7
(0.2,0.8)	73.9
(0.3,0.7)	75.5
(0.4,0.6)	76.9
(0.5,0.5)	75.5

从表 2 中可以得到：1)“batch hard”三元组采样方案并不适用于基于非线性度量的三元卷积神经网络的模型，模型性能只能达到 56.2%。其主要原因是基于非线性度量的“batch hard”方法会采样出许多对训练不利的三元组，导致模型收敛效果差。2)基于“limited batch hard”三元组采样规则，模型可以更好地收敛，随着“limited batch hard”的加入以及限制逐渐加强，模型的识别能力从 72.7%开始提升，当限制正样本图像对关系值大于或等于 0.4，且负样本图像对关系值小于或等于 0.6 时，性能达到 76.9%，相比“batch hard”性能提升了 20.7%。“limited batch hard”利用正负样本图像对的关系值来限制三元组的选择，可以采样出更加适合融合关系网络和三元卷积神经网络的三元组，最终相比“batch hard”挑选规则，本文模型的训练更加稳定，准确率更高。

4.3.4 模型初始化分析

初始化参数对于本文所提模型十分重要，我们在 VIREO Food-172 数据集上探究了不同的初始化方案对模型性能的影响。为了分析特征嵌入网络与关系网络的不同初始方案，我们使用“微调”表示初始化方案特定的设置，具体含义如下：对于特征嵌入网络，“否”代表使用 ImageNet 预训练模型初始化，“是”代表由数据集的训练集预训练模型初始化；对于关系网络，“否”代表随机初始化，“是”代表由训练集预训练的模型初始化。实验结果证明了所提出的模型参数初始方案的重要性。从表 3 可以得到：1)关系网络的初始化对于模型的训练是至关重要的。如果采用随机初始化，无论特征嵌入网络采用何种初始化方案，模型在 3 个数据集上都将无法收敛。关系网络是采样三元组的度量方法，如果非线性网络使用随机初始化，则挑选出的三元组对于三元神经网络的训练是不具有参考性的。因此，我们需要对关系网络进行一个较好初始化。2)特征嵌入网络的初始化对实验性能有影响，相较 ImageNet 初始化，基于训练集的初始化可以获得更好的性能，在 3 个数据集上性能大约提升 3.0%。其主要原因是特征嵌入网络为三元组选择提供特征表示，特征表示的区分程度也决定采样的三元组是否适合训练。因此，最终采用基于训练集获得的分类模型参数进行模型的初始化，并取得性能提升。

表 3 不同参数初始方案的实验结果

Table 3 Experimental results of different parameter initialization scheme

（单位：%）

微调（特征嵌入网络）	微调（关系网络）	Food-101	VIREO Food-172	Chinese-FoodNet
否	否	—	—	—
是	否	—	—	—
否	是	53.2	73.3	60.9
是	是	56.3	76.9	66.1

结束语 本文提出了一个融合非三元卷积神经网络和关系网络的小样本食品识别模型，该模型可以学习到图像类内和类间的区分信息，同时关系网络作为可学习的非线性度量函数，可以更好地指导图像特征嵌入网络的学习，并更好地契合数据集的特性。为训练该模型，我们进一步提出了一个新的基于关系网络的三元组的采样规则“limited batch hard”，新的采样方法可以有效地筛选、去除对训练不利的三元组。我们在不同的食品数据集上进行实验，所有的实验结果证明了所提方法的有效性。同时通过对比现有流行的样本学习方法，本文的模型取得了最好的性能。

目前，仍有许多问题值得进一步探究：1)可以探索食品数据集中更加丰富的信息提升性能。例如，食品属性信息、地理位置信息和食谱信息等。2)由于本文方法具有普适性，因此我们希望可以将此方法应用到更多的领域中，例如 Omniglot[10] 和 MiniImageNet[11] 等。

参 考 文 献

[1] BOSSARD L, GUILLAUMIN M, VANGOOL L. Food-101-mining discriminative components with random forests[C]// European Conference on Computer Vision. 2014: 446-461.

[2] AO S, LING C X. Adapting new categories for food recognition with deep representation[C]// IEEE International Conference on Data Mining Workshop. 2015: 1196-1203.

[3] HERRANZ L, JIANG S, XU R. Modeling restaurant context for food recognition[J]. IEEE Transactions on Multimedia, 2017, 19(2): 430-440.

[4] AIZAWA K, MARUYAMA Y, LI H, et al. Food balance estimation by using personal dietary tendencies in a multimedia foodlog[J]. IEEE Transactions on Multimedia, 2013, 15(8): 2176-2185.

[5] ZHENG J, WANG Z J, ZHU C. Food image recognition via superpixel based low-level and mid-level distance coding for smart home applications[J]. Sustainability, 2017, 9(5): 856.

[6] BOLANOS M, FERRA A, RADEVA P. Food ingredients recognition through multi-label learning[C]// International Conference on Image Analysis and Processing. 2017: 394-402.

[7] ZHANG N, DONAHUE J, GIRSHICK R, et al. Part-based r-cnns for fine-grained category detection[C]// European Conference on Computer Vision. 2014: 834-849.

[8] CHRISTODOULIDIS S, ANTHIMOPOULOS M, MOUGIAKAKOU S. Food recognition for dietary assessment using deep convolutional neural networks[C]// International Conference on

Image Analysis and Processing. 2015:58-465.
[9] MARTINEL,NIKI,FORESTI G,et al. Wide-Slice Residual Networks for Food Recognition[C]// IEEE Winter Conference on Applications of Computer Vision IEEE Computer Society. 2018:567-576.
[10] KOCH G,ZEMEL R,SALAKHUTDINOV R. Siamese neural networks for one-shot image recognition[C]// International Conference on Machine Learning. 2015.
[11] VINYALS O,BLUNDELL C,LILLICRAP T,et al. Matching networks for one shot learning[C]// Advances in Neural Information Processing Systems. 2016:3630-3638.
[12] SUNG F,YANG Y,ZHANG L,et al. Learning to compare:Relation network for few-shot learning[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2017.
[13] FINN C,ABBEEL P,LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[M]. arXiv:1703. 03400, 2017.
[14] ANDRYCHOWIEZ M,DENIL M,GOMEZ S,et al. Learning to learn by gradient descent by gradient descent[C]// Advances in Neural Information Processing Systems. 2016:3981-3989.
[15] CEALLE S,MANINIS K,PONTTUEST J,et al. One-Shot Video Object Segmentation[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE,2017.
[16] HOFFE E,AILON N. Deep metric learning using triplet network[M]. In International Workshop on Similarity-Based Pattern Recognition,2015.
[17] HERRMANS A,BEYER L,LEIBE B. In defense of the triplet loss for person re-identification[J]. arXiv:1703. 07737,2017.
[18] GENG M,WANG Y,XIANG T,et al. Deep transfer learning for person re-identification[J]. arXiv:1611. 05244,2016.
[19] LI Y,LI Y,YAN H. Deep joint discriminative learning for vehicle re-identification and retrieval[C]// IEEE International Conference on Image Processing. IEEE,2017:395-399.
[20] CHEN J,NGO C W. Deep-based ingredient recognition for cooking recipe retrieval[C]// Proceedings of the ACM International Conference on Multimedia. 2016:32-41.
[21] CHEN X,ZHOU H,ZHU Y,et al. Chinesefoodnet:A largescale image dataset for chinese food recognition[J]. arXiv: 1705. 02743,2017.
[22] MIN W Q,JIANG S Q,LIU L H,et al. A Survey on food computing[J/OL]. https://arxiv. org/abs/1808. 07202?context = cs. mm
[23] KAWANO Y,YANAI K. Food image recognition with deep convolutional features[C]// Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing:Adjunct Publication. 2014:589-593.
[24] KAGAYA H,AIZAWA K,OGAWA M. Food detection and recognition using convolutional neural network[C]// Proceedings of the ACM International Conference on Multimedia. 2014:1085-1088.
[25] XU R,HERRANZ L,JIANG S Q. Geolocalized Modeling for Dish Recognition[J]. IEEE Transactions on Multimedia,2015, 17(8):1187-1199.
[26] MIN W Q,JIANG S Q,SANG J T,et al. Being a super cook:Joint food attributes and multimodal content modeling for recipe retrieval and exploration[J]. IEEE Transactions on Multimedia, 2017(5):1100-1113.
[27] MIN W Q,BAO B K,MEI S H,et al. You are what you eat:Exploring rich recipe information for cross-region food analysis[J]. IEEE Transactions on Multimedia,2017,20(4):950-964.
[28] WANG H,MIN W,LI X,et al. Where and what to eat:Simultaneous restaurant and dish recognition from food image[C]// Pacific Rim Conference on Multimedia. 2016:520-528.
[29] MEI S H,MIN W Q,LIU L H. Faster R-CNN based food image retrieval and classification [J]. Journal of Nanjing University of Information Science & Technology (Natural Science Edition), 2017,9(6):635-641.
[30] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409. 1556, 2014.
[31] KINGMA D,BA J. Adam:A method for stochastic optimization [C]// arXiv:1412. 6980. 2014.
[32] MENG Y,GUO Y. Deep Triplet Ranking Networks for One-Shot Recognition[J]. arXiv:1804. 07275,2018.

LV Yong-qiang, born in 1992, postgraduate. His main research interests include deep learning, computer vision and machine learning.

DUAN Hua, born in 1976, Ph.D, professor. Her main research interests include Petri nets, process mining and machine learning.