北京航空航天大学学报
Journal of Beijing University of Aeronautics and Astronautics
ISSN 1001-5965,CN 11-2625/V

《北京航空航天大学学报》网络首发论文

题目：基于噪声柯西分布的社交图像标签优化与标注
作者：练连荣，项欣光
DOI：10.13700/j.bh.1001-5965.2020.0454
收稿日期：2020-08-24
网络首发日期：2020-10-16
引用格式：练连荣，项欣光．基于噪声柯西分布的社交图像标签优化与标注．北京航空航天大学学报. https://doi.org/10.13700/j.bh.1001-5965.2020.0454

基于噪声柯西分布的社交图像标签优化与标注

练连荣[1]，项欣光[1]

（1. 南京理工大学 计算机科学与工程学院，南京 210094）
*通信作者 项欣光 E-mail： xgxiang@njust.edu.cn

摘要 随着社交网络的快速发展，带有用户提供标签的社交网络图像呈现爆炸式增长。但是用户提供的标签是不准确的，存在很多不相关以及错误的标签。这势必会增加相关多媒体任务的困难。标签噪声虽然是杂乱无序的，但是依然符合一定的概率分布。之前方法大多采用高斯分布来拟合噪声，但是高斯分布对大噪声比较敏感。鉴于此，本文采用对各种噪声都具有鲁棒性的柯西分布拟合噪声。进而，本文提出了一个基于噪声柯西分布的非负低秩深度学习模型（CDNL），通过柯西分布建模标签噪声来获得理想标签，并利用深度神经网络模块学习视觉特征和理想标签之间的内在联系，来得到图像对应的正确标签，从而大幅提高社交网络图像的标签准确率。本模型不仅可以修正错误标签，补充缺失标签，也可以对新图像进行标注。本论文在两个公开的社交网络图像数据集上进行了验证，并且与一些最新的相关工作进行了对比，证实了该方法的有效性。

关键词 社交标签；柯西分布；深度神经网络；图像标注；矩阵分解
中图分类号 V221+.3；TB553 DOI: 10.13700/j.bh.1001-5965.2020.0454

Social Image Tag Refinement and Assignment Based on Noise Cauchy Distribution

Lian Lianrong[1] , XIANG Xinguang[1]

（1. School of Computer Science and Engineering, Nanjing University of Science and Technology,Nanjing 210094, China ）
*XIANG Xinguang, E-mail: xgxiang@njust.edu.cn

Abstract With the rapid development of social networks, images with social tags have also exploded. However, these tags are usually inaccurate and irrelevant which will make it harder for the relevant multimedia tasks. Although label noise is chaotic and disordered, it still conforms to a certain probability distribution. Most of the current methods use Gaussian distribution to fit the noise, but Gaussian is very sensitive to large noise. So we use the Cauchy distribution to fit the noise which is robust to various noises. In this paper, we propose a weakly-supervised non-negative low-rank deep learning model(CDNL). Which build the noise model by Cauchy distribution to obtain the ideal label and use CNN to reveal the intrinsic connection between the visual features of the image and the labels. The proposed method can not only correct wrong labels and add missing labels, but also tags new images. Experiments are conducted on two multi-label public datasets. Compared with some of the latest related work, qualitative and quantitative results show the effectiveness of the proposed method.

Key words social tags; Cauchy distribution; deep neural networks; image annotation; matrix factorization

近年来随着社交网络的蓬勃发展，海量的社交网络图像在被网络用户共享和浏览。海量的图像数据使得精确检索变得困难，故现在迫切需要有效的图像检索技术。基于标签的图像检索是通过建立图像和标签之间的语义关系来进行的。图像标注对图像检索十分重要。目前，用户提供的标签虽然可以在一定程度上描述视觉内容信息，但是这些标签是不准确的。如工作[1]中所述，用户提供的标签只有一半可以描述图像的视觉内容。这是因为现实中社交网络标签通常是不完整、不准确的，甚至有非常

收稿日期：2020-08-24
作者简介：项欣光 男，副教授，硕士生导师。主要研究方向：视频处理，压缩与通讯，智能媒体分析、图像处理。练连荣 男，硕士研究生。主要研究方向：社交媒体多标签分类。
网络首发时间：2020-10-16 16:00:45 网络首发地址：https://kns.cnki.net/kcms/detail/11.2625.v.20201016.1445.001.html

大比例的图像是没有标签的（在文献[2]中 MIRFlickr 数据集中有超过 50%的图片没有标签）。也就是说用户提供的标签是弱监督的。而这会增加相关多媒体任务的困难，所以通过学习图像视觉信息和标签语义之间的内在联系来提高社交图像的标签质量是非常必要的。

对于社交网络图像的重标注问题，之前的工作提出了多种解决方法，例如文献[3]采用矩阵分解来最小化噪声，从而学习图像-标签的内在关系。另外工作[4]验证了重标注标签矩阵的低秩性，并且考虑了图像视觉特征的一致性和标签之间的相关性。基于工作[4]中的低秩非负模型，工作[5][6][7]通过引入两个潜在因子矩阵来分离优化函数，更好的减小重标注标签和观测标签之间的差别，并得到更理想的图像-标签关系模型。上述工作均验证了低秩非负模型在社交图像重标注任务中的有效性。

在上述工作[3,4,5]中都采用了矩阵分解的方法，并且基于模型泛化能力的考虑，通常采用平方损失函数作为目标函数。然而这其中隐含的基本假设是社交标签中的噪声是服从高斯分布的[8]。由于中心极限定理[9]的特点，高斯概率密度函数被广泛应用于信号处理和图像分析等领域中。高斯分布可以很好的拟合最常出现的白噪声。但是在现实中，数据噪声的内在概率分布是未知的，可能存在各种类型的噪声。然而高斯分布能够很好的拟合小噪声的分布情况，而对大噪声比较敏感[10]。鉴于上述情况，本文选用对大噪声更加鲁棒的柯西分布建模标签噪声，并由于噪声的稀疏性，采用l_1范数来优化噪声矩阵。

另外在图像标注任务中，图像特征对模型训练具有显著意义，具有更深语义的图像特征能显著提升训练效果。在工作[3][5][11]中都采用了传统的特征提取方法（如 Gist、SIFT、HOG），这会导致深层语义的丢失。鉴于此，本文根据工作[12]提出利用预训练好的 Resnet50 网络来提取视觉特征。基于提出的视觉特征和观测标签，本文采用三层 CNN 网络学习视觉特征与高层语义之间的映射关系。此外，为了更快的训练模型，会利用特征矩阵和观测标签矩阵对模型进行预训练。

鉴于此，本文提出了一种基于柯西噪声分布的弱监督非负低秩深度模型(CDNL)，来同时解决错误标签重新标注，以及新图像标注问题。非负低秩模型主要用于对理想标签的优化和对噪声的抑制，通过引入两个潜在因子分离优化目标。柯西分布能够更好的拟合标签噪声，通过优化损失函数，减少理想标签矩阵和观测标签矩阵的差异。另外，为了更好的学习图像和理想标签的内在映射关系，本文选用 Resnet50 网络来提取深层图像特征，以及 CNN 网络来学习特征矩阵和标签矩阵的内在关系。本实验基于两个广泛使用的社交图像数据集，并且和一些最新的先进工作进行对比，来证实本方法的有效性。

本文贡献在于提出了一种基于噪声柯西分布的弱监督非负低秩深度学习框架，创新性地利用柯西分布模型建模标签噪声，学习得到视觉特征和标签语义之间更深层次的内在关系。通过建立不同尺度的柯西分布模型，模拟标签噪声，选择出最拟合标签噪声的柯西模型，从而更好地减少理想标签矩阵和观测标签矩阵之间的差异。另外本文首先采用预训练的 Resnet50 模型来提取特征，并利用 CNN 网络来学习特征与标签之间的映射关系。本模型不仅能够细化、补全、重标注社交图像，而且可以标注新图像。定性定量的结果证明了我们方法优于其他方法。

1 相关工作

在图像标注领域，许多先前的工作致力于通过学习图像-标签关系模型来提高图像标签质量。其中工作[13][14][15]是基于监督模型上的，但是随着用户提供的社交图像呈现爆炸式增加，弱监督图像标注与优化成为图像标注领域的难点，大量工作利用传统方法（如[6][11][16][17][18]）和深度学习方法（[3][5][19][20][21][22][23][24][25] [26]）来解决这个问题。传统方法通过学习标签之间的语义相关性来探索用户提供的标签语义信息。比如在[6]中，通过邻居投票策略来研究图像标签之间的相关性。在[11]中通过删除不准确的标签和增加相关标签，来探究视觉相似性和语义相关性之间的一致性。在[17]中通过低阶矩阵分解来解决标签优化问题。在[18]中通过结合低秩矩阵复原和最大似然估计，来对缺失的标签进行补充，对噪声标签进行细化。然而上述工作在图像视觉特征和学习到的低秩标签空间的联系上缺乏有效性，使得准确性受到限制。

而深度学习方法的出现在一定程度上解决了这一问题。工作[19]证明了深度学习方法在图像标注领域是十分有效的，近几年深度学习方法也广泛应用于图像标注任务，如在[20]中认为修正、补全、重标注社会标签的过程，其本质是广义协同过滤的过程。广义协同方法即：相似的图像应该具有相似的标签，具有相同标签的图像也应该具有相似的内容。因此应用协同过滤的方法来处理社会标签是可行的，并提出了 CCA 模型来对标签进行标注。在[3][5]中则为了优化非负低秩矩阵模型，通过引入潜在因子来分离优化函数，来达到分别优化理想标签和噪声标签的目的，并且利用了深度学习模型来训练图像-标签的内在关系。其中工作[5]将基于社会标签的图像标注问题表述为：通过获得低秩的重标注矩阵和稀疏的误差矩阵，来合成观测标签矩阵的问题（由于图像和标签之间的语义相关性，重标注矩阵应当为低秩矩阵，而标注误差不常出现，所以误差矩阵应该是稀疏的），并且对该问题进行建模优化。在[23]中，通过假设噪声为高斯分布，并结合图像相似性，标签一致性以及矩阵分解模型，来优化基于损失平方和的目标函数。工作[24]利用图卷积网络在部分标签上进行学习，并允许不需要附加标签的组级联接，借此来学习图像-标签关系模型。在工作[25]中，提出利用语义图嵌入的跨模态注意力机制来学习语义标签嵌入，并显式的利用了标签之间的关系，在学习标签嵌入的指导下，生成新颖的跨模态注意力图。在工作[26]中则在非负低秩矩阵分解的基础上，引入深度学习方法学习标签语义图。上述工作均验证了深度学习能够有效的建立起图像特征和标签语义之间的联系，并且工作[3][5][23][26]验证了矩阵分解方法优化社交标签的有效性。然而上述工作默认噪声符合高斯分布，而实际数据的内在概率分布是未知的，不一定符合高斯分布，所以需要对噪声建模。

在真实噪声环境中，噪音源往往是多样的。假设噪音是由多个不同概率分布的随机变量加合而成，并且每一个随机变量都是独立的，那么根据中心极限定理[9]，噪音分布随着噪音源数量的上升趋近于高斯分布[10]。虽然高斯分布能有效解决小噪声但是对于大噪声过于敏感，而社交图像如工作[2]所述属于大噪声。柯西分布的重尾特性可以有效的建模大噪声，并且柯西分布在峰值处平滑，使得其对密集噪声也有很好的效果[27]。所以本文选用柯西分布来拟合标签噪声。

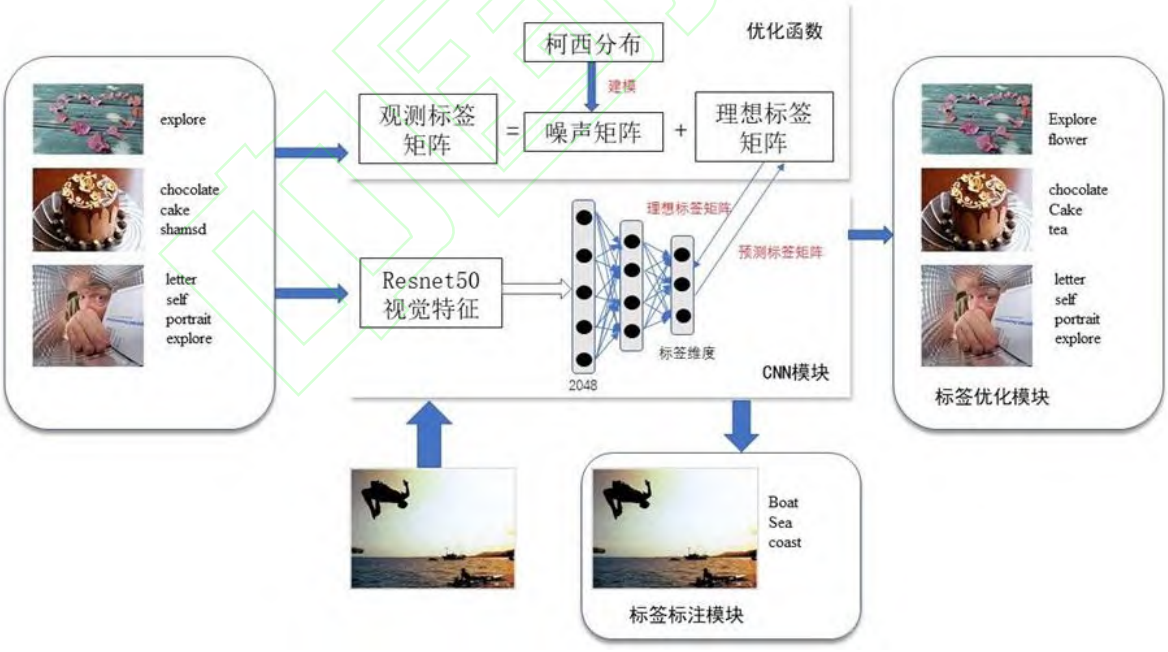

图 1　基于噪声柯西分布的社交网络图像标注和重标注

基于上述工作，本文提出在非负低秩矩阵模型的基础上，基于噪声稀疏性原则，利用柯西分布来对噪声建模（图 1）。并选择 Rsenet50 网络提取更深层次的视觉特征，为了更好的联系视觉特征和标签空间，本文还利用 CNN 网络来训练图像和理想标签矩阵模型，并在训练模型前，对 CNN 网络预训练来减少预测标签和理想标签间的差异。

2 柯西噪声模型

在本节中，将介绍基于柯西噪声分布的弱监督非负低秩深度模型(CDNL)（如图一），着重阐述本文动机，并对优化函数进行推导。

2.1 动机

对于图像标签标注问题，最重要的是要揭示视觉内容和语义标签的内在联系。而对于社交网络图像而言，用户通常会提供一些标签用来标注图像。而这些标签形成的语义空间是可以由真实标签空间中的显式标签子集来近似，而且用户倾向于选择语义相关的标签来对图像进行标记。因此，社交网络图像的标签-图像关联矩阵本质是低秩的。此外，标注过程中不常出现错误，所以噪声矩阵是稀疏的。这是符合常理的。因此，本文采用低秩非负模型来解决社交图像标注问题。为了更好的解决标签噪声分布问题，本文选择对各种噪声更加鲁棒的柯西概率分布拟合噪声，由此建立基于噪声柯西分布的低秩非负模型（CDNL）。图像的视觉特征对图像-标签学习至关重要，因此本文选择 Resnet50 来提取深层视觉特征。另外，尽管观测标签矩阵仍然存在很多不相关以及错误的标签，图像的视觉特征和标签仍然是紧密联系的，所以本文采用 CNN 框架来学习图像和标签之间的映射关系。

2.2 模型

本文中定义斜体大写字母(如A)为矩阵，A_{ij} 表示矩阵A的第$(\mathrm{i},\ \mathrm{j})$元素，矩阵$A$的核范数为$\|A\|_*$，矩阵$A$?$l_1$????$A$?$_1 = ?_{i=1}^{m}?_{j=1}^{n}|A_{i,j}|$，而矩阵$A$的 F 范数为?$A$?$_F^2 = ?_{i=1}^{m}?_{j=1}^{n}A_{i,j}^2$。对于图像标注问题，图像数据集包括$n$幅图像和$m$个用户标签，每幅图像对应若干个标签构成二值矩阵$F$，即观测标签矩阵。本文定义理想标签矩阵为$Y$。定义特征矩阵为$X$，由数据集每幅图像对应的特征向量$x_i$构成。在低秩框架下，最重要的是优化噪声矩阵$E$、理想标签矩阵$Y$以及标签预测对应的损失函数。观测标签矩阵由理想标签矩阵和噪声标签矩阵组成：

$$F = Y + E \tag{1}$$

根据文献[28]中的工作，用 rank(Y)（Y的秩）来衡量标签矩阵Y的低秩性，$S(E)$（由不同的噪声优化目标决定）来衡量噪声矩阵E的稀疏性，$\mathrm{loss}\left(Y, W_{\mathrm{g}}(X)\right)$来衡量标签预测的损失，其中$W_g(X)$是 CNN 网络预测标签，最后为了防止过拟合问题，再引入正则化项$\Omega(\theta)$，并且引入超参数λ_1，λ_2以及λ_3来构成模型的优化函数：

$$\begin{aligned}&\min_{Y\geq 0,\ W} rank(Y)+\lambda_1 S(E)+\lambda_2 loss\left(Y,\ W_{\mathrm{g}}(X)\right)\\&+\lambda_3\Omega(\theta)\end{aligned} \tag{2}$$

对于上述优化函数，利用$\|Y\|_*$来衡量Y的秩，则$rank(Y) = ?Y?_*$，而对于$S(E)$，假设噪声矩阵E符合柯西分布：

$$\mathrm{p}(E_{ij}) = \frac{b}{\pi}\frac{1}{b^2 + E_{ij}^2} \tag{3}$$

则对于$S(E)$的优化为：

$$\begin{aligned}&\min \sum_{i=1}^{\mathrm{m}}\sum_{j=1}^{\mathrm{n}}\ln\left(b^2+\left(F_{ij}-Y_{ij}\right)^2\right)\\&=\left\|\ln\left(b^2+(F-Y)^2\right)\right\|_1\end{aligned} \tag{4}$$

对于预测误差采用$loss(x,y) = \frac{1}{2}(x-y)^2$的损失函数，则最后优化目标为：

$$\begin{aligned}&\min_{Y\geq 0,W}\|Y\|_* + \lambda_1\left\|\ln\left(b^2+(F-Y)^2\right)\right\|_1 + \tfrac{\lambda_2}{2}\left\|Y-W_g(X)\right\|_F^2\\&+\lambda_3\Omega(\theta)\end{aligned} \tag{5}$$

对上述优化问题进行优化，根据文献[3][28]，引入两个辅助因子Y_1, Y_2分离优化问题，得到以下优

化问题：

$$\begin{aligned}&\min_{Y\geq 0,W}\|Y_1\|_*+\lambda_1\left\|\ln\left(b^2+\left(F-Y_2\right)^2\right)\right\|_1+\tfrac{\lambda_2}{2}\left\|Y-W_g\left(X\right)\right\|_F^2\\&+\lambda_3\Omega\left(\theta\right)\\&\text{s.t. } Y_1=Y,Y_2=Y,Y_2\geq 0\end{aligned}\tag{6}$$

再根据非精确增广拉格朗日方法[28]，对应于上式中的Y_1,Y_2，引入Z_1,Z_2得到增广拉格朗日方程为：

$$\begin{aligned}\xi=&\|Y_1\|_*+\lambda_1\left\|\ln\left(b^2+\left(F-Y_2\right)^2\right)\right\|_1+\tfrac{\lambda_2}{2}\left\|Y-W_g\left(X\right)\right\|_F^2\\&+\tfrac{\eta}{2}\left(\left\|Y-Y_1+\tfrac{1}{\eta}Z_1\right\|_F^2+\left\|Y-Y_2+\tfrac{1}{\eta}Z_2\right\|_F^2\right)-\tfrac{1}{2\eta}\left(\left\|Z_1\right\|_F^2+\left\|Z_2\right\|_F^2\right)\end{aligned}\tag{7}$$

通过对上述优化函数求解偏导，得到Y，Y_1，Y_2的偏导。如下所示：

$$\begin{aligned}&\arg\min_{Y}\tfrac{\lambda_2}{2}\left\|Y-W_g\left(X\right)\right\|_F^2+\tfrac{\eta}{2}\left(\left\|Y-Y_1+\tfrac{1}{\eta}Z_1\right\|_F^2\right.\\&\left.+\left\|Y-Y_2+\tfrac{1}{\eta}Z_2\right\|_F^2\right)\\&\arg\min_{Y_1}\|Y_1\|_*+\tfrac{\eta}{2}\left\|Y-Y_1+\tfrac{1}{\eta}Z_1\right\|_F^2\\&\arg\min_{Y_2}\lambda_1\left\|\ln\left(b^2+\left(F-Y_2\right)^2\right)\right\|_1+\tfrac{\eta}{2}\left\|Y-Y_2+\tfrac{1}{\eta}Z_2\right\|_F^2\end{aligned}$$

得到Y，Y_1，Y_2的更新公式：

$$Y=\tfrac{1}{2\eta+\lambda_2}\left(\lambda_2+W_{\mathrm{g}}\left(X\right)-Z_1-Z_2+\eta Y_1+\eta Y_2\right)\tag{8}$$

$$Y_1=\Gamma_{\frac{1}{\eta}}\left(Y+\tfrac{1}{\eta}Z_1\right)\tag{9}$$

$$Y_2=F-soft\left(F-\tfrac{1}{\eta}Z_2-Y,\tfrac{\lambda_1}{\eta},\sqrt{1-b^2}\right)\tag{10}$$

在上述公式中$\Gamma_\delta=Usoft\left(\Lambda,\ \delta,\ \tau\right)V^T$，$soft\left(A_{ij},\ \delta,\ \tau\right)=sign\left(A_{ij}\right)\max\left(\left|A_{ij}\right|-\delta,\ \tau\right)$，$A=U\Lambda V^{\mathrm{T}}$(其中$\Lambda$是$A$的 SVD 分解)。

综上所述，本文首先提出了优化目标函数（公式（2）），并对三部分优化目标引入不同的损失函数。根据工作[28]，通过非精确增广拉格朗日方法来求解目标函数（公式（6）），并对目标函数中的变量求解偏导数，来得到变量Y，Y_1，Y_2的迭代过程，而神经网络参数W则根据 CNN 网络进行更新。根据工作[28]，非精确增广拉格朗日方法可以有效的保证该算法的收敛性。

本文通过输入预测标签矩阵$W_g(X)$，经过上述的优化过程，可以计算得到理想的标签矩阵Y。再将视觉特征X和理想的标签矩阵Y输入到 CNN 网络中，学习图像-标签关系，并得到新的观测标签矩阵$W_g(X)$，不断迭代上述过程直至收敛。再利用训练得到图像-标签关系模型，来对已有图像重标注，并对新图像标注。下面介绍本论文 CDNL 模型的算法流程，其中本论文 CDNL 模型对数据集 MIRFlickr 采用四层网络结构，对 NUS-WIDE 采用三层网络结构，但这不会影响优化模块的结果。

CDNL 算法：

输入：深层视觉特征X以及观测标签矩阵F。

预训练：将X，F输入 CNN 网络，设置 Sigmoid 函数为激活函数，二分类交叉熵函数（Binary Cross Entropy）为损失函数，学习率为 0.0001，指数衰减率为（0.9, 0.999），预先训练 CNN。

准备：载入预训练 CNN，随机正态初始化 Y_1, Y_2, Z_1, Z_2

训练:repeat:

$$\eta = 0.1, \rho = 1.1$$

根据公式(8)(9)(10)更新Y, Y_1, Y_2

$$Z_1 = Z_1 + \eta\left(Y - Y_1\right)$$

$$Z_2 = Z_2 + \eta\left(Y - Y_2\right)$$

$$\eta = \eta * \rho$$

CNN 模型更新$W_g(X)$（参数同上）

Until 模型收敛

输出：理想标签矩阵Y，模型参数W，评价数据

使用：利用训练的模型，来重标注训练集，并对测试集中新图片进行标注

3 实验

在这部分将会详细介绍本方法的实验流程，并且对实验结果进行分析。

本实验主要基于两个公开的社交网络图像数据集：MIRFlickr 和 NUS-WIDE，这两个数据集都在社交网络图像理解和检索任务上得到广泛的应用。MIRFlickr[29]数据集有 25000 幅图像、386 个用户标签，去掉出现少于 50 次的标签，得到 457 个实验标签，另外有 18 个真实有效标注用于评价性能。NUS-WIDE[30]数据集有 269648 幅图像和 5018 个用户标签，挑选出现频率最高的 1000 个标签作为实验标签，161777 幅图像作为实验数据集。另外有 81 个真实有效标注用于评价。实验中所用数据集的信息如表 1 所示。

表1 MIRFlickr和NUS-WIDE实验数据

	MIRFlickr	NUS-WIDE
图片数量	25000	269627
标签数量	457	1000
真实标签数量	18	81
每次训练集图片数量	12500	160000

为了验证模型的稳定性，我们从数据集中随机抽取图像作为训练集，并用剩下的图像作为测试集，并且重复实验 5 次。对于 MIRFlickr 每次抽取 12500 幅图像作为训练集，剩下 12500 幅图像作为测试集。对于 NUS-WIDE 每次抽取 160000 幅图像作为训练集，剩余图像作为测试集。

由于传统方法只能提取图像的底层视觉信息。为了得到图像更深的视觉内容，本实验采用 Resnet50 残差网络来提取 2048 维的图像特征。为了使预测标签更接近于理想标签，训练开始时会将深层视觉特征X以及观测标签矩阵F输入 CNN 来预先训练网络。这样能够更有效的训练 CDNL 模型。进而，不断优化理想标签矩阵Y，和网络模型参数W。最后利用已训练好的模型来重标注训练集，并标注测试集新图像。

本文对于目标函数中的超参数λ_1，λ_2以及λ_3，从｛0.001,0.01,0.05,0.1,0.5,1,5,10,100,1000｝中依次选择实验，对比实验结果。其中 MIRFlickr 数据集上λ_1为 0.1，λ_2为 0.2 以及λ_3为 0.0001，NUS-WIDE 上λ_1为 0.7，λ_2为 0.5 以及λ_3为 0.0001 时最优。

本文采用 MicroAUC，MacroAUC 以及平均精度均值（mAP）这三个指标来对模型进行评价。

将 CDNL 模型和前人的工作进行对比，包括 LSCCA[7]，CCA-CNN[31]，TCCA[18]，DMF[32]，

MPMF[33]，DNMF[34]，WDMF[3]，WDNL[5]，DCE[23]等模型，其中 LSCCA，TCCA 属于传统方法，CCA-CNN 属于利用深度学习方法学习图像-标签关系，DMF，MPMF 以及 DNMF 属于矩阵分解方法，而 WDMF，WDNL 和 DCE 均在标签矩阵分解的基础上学习了图像视觉空间和标签语义空间的关系。

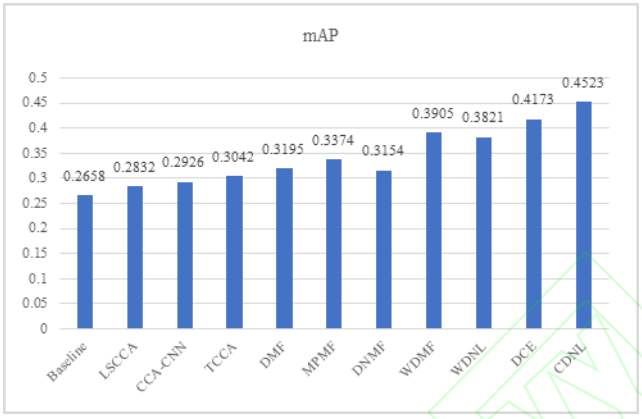

图 2 MIRFlick 的重标注 mAP

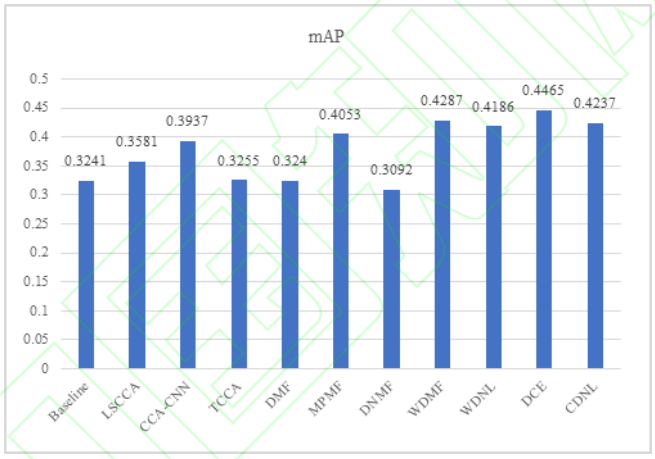

图 3 NUS-WIDE 的重标注 mA

表2 在MIRFlickr和NUS-WIDE进行标签重标注的实验结果（平均microauc/macroauc±标准偏差值）

Method	MIRFlickr		NUS-WIDE	
	MicroAUC	MacroAUC	MicroAUC	MacroAUC
Baseline	0.558	0.587	0.623	0.754
LSCCA	0.594±0.006	0.586±0.004	0.732±0.002	0.632±0.003
CCA-CNN	0.644±0.006	0.631±0.003	0.675±0.005	0.743±0.007
TCCA	0.643±0.006	0.632±0.004	0.768±0.007	0.675±0.008
DMF	0.639±0.002	0.628±0.002	0.751±0.005	0.739±0.004
MPMF	0.634±0.004	0.607±0.002	0.782±0.002	0.681±0.005
DNMF	0.624±0.005	0.621±0.006	0.759±0.009	0.665±0.003
WDMF	0.704±0.005	0.678±0.006	0.805±0.004	0.775±0.006
WDNL	0.685±0.003	0.671±0.003	0.789±0.006	0.762±0.006
DCE	0.732±0.003	0.718±0.004	0.825±**0.004**	0.797±0.003
Ours	0.745±**0.004**	0.775±**0.006**	0.774±0.009	0.831±**0.005**

表3 在MIRFlickr和NUS-WIDE进行新图像标签标注的实验结果（平均microauc/macroauc±标准偏差值）

Method	MIRFlickr		NUS-WIDE	
	MicroAUC	MacroAUC	MicroAUC	MacroAUC
LSCCA	0.585±0.006	0.562±0.004	0.681±0.002	0.599±0.003
CCA-CNN	0.642±0.005	0.627±0.002	0.617±0.004	0.641±0.003
TCCA	0.610±0.005	0.597±0.006	0.727±0.009	0.625±0.006
DMF	0.635±0.002	0.623±0.003	0.737±0.004	0.632±0.004
MPMF	0.617±0.004	0.596±0.002	0.742±0.002	0.635±0.005
DNMF	0.619±0.005	0.601±0.006	0.699±0.009	0.618±0.003
WDMF	0.661±0.007	0.646±0.004	0.768±0.005	0.675±0.007
WDNL	0.665±0.004	0.652±0.005	0.758±0.004	0.671±0.007
DCE	0.693±0.005	0.667±0.004	0.787±**0.006**	0.746±0.004
Ours	0.715±**0.006**	0.735±**0.005**	0.761±0.008	0.794±**0.007**

本文提出的 CDNL 模型将从图像重标注和新图像标注两个方面进行评价，从图 2 可以看出 CDNL 模型在 MIRFlickr 上效果显著，远高于当前主流方法，从图 3 可以看出 CDNL 模型在 NUS-WIDE 上效果也有提升，仅仅稍差于 DCE。而从表 2 和表 3 可以发现论文提出的模型在两个方面的两个指标上均提升明显，相较于其他方法在 MIRFlickr 上有显著优势，在 NUS-WIDE 上也有进步。其中相较于 LSCCA，TCCA 等传统方法，矩阵分解（如 DMF，MPMF，DNMF）有效的提升了标签准确率，但是由于没有考虑图像特征和标签语义的关系，效果明显不如 WDMF，WDNL 和 DCE，而本文方法在引入噪声柯西分布拟合后，对比 WDMF，WDNL 方法在两个数据集上的效果均有所提升。提升幅度有大有小，这是由于数据集噪声分布是未知的，高斯分布或柯西分布均没有办法完全拟合噪声分布，但是柯西分布相对高斯分布依然有一定优势。由此我们可以看出柯西分布对噪声的拟合，Resnet50 对图像深层语义内容的提取以及 CNN 的训练都是卓有成效的，CDNL 模型对比其他主流方法有很大的提升。

而对于噪声柯西分布的具体形式，本实验也做了相关对比工作。在柯西分布中，尺度参数 b 代表了分布中最大值一半处的宽度，并代表着噪声分布的状态，通过调整 b 值分析数据集中噪声的分布特征和状态，在两个数据集上的实验结果如表 4 所示。可以看到 MIRFlickr 更加拟合尺度参数 b 为 0.8 的柯西分布，而 NUS-WIDE 更加拟合尺度参数为 0.6 的柯西分布。

表4 不同尺度参数b对数据集mAP性能的影响

	0.2	0.4	0.6	0.8
MIRFlickr	0.417	0.423	0.425	0.452
NUS-WIDE	0.401	0.415	0.423	0.403

从上面的实验数据可以看出我们提出的模型在数据集 MIRFlickr 上效果远好于前人的工作，并且在 b 为 0.8 的效果最为优异，说明 MIRFlickr 的噪声非常符合柯西分布，且基于噪声柯西分布的深度模型是十分有效的。而在数据集 NUS-WIDE 上，我们的结果稍逊于 DCE 方法，但是也有所提升，并在 b 为 0.6 时得到最好效果，说明 NUS-WIDE 噪声分布不是十分拟合柯西分布，但是柯西分布由于其对大多数噪声的适用性，仍然很好的提升了模型的鲁棒性。

4 结论

本文提出了一种基于噪声柯西分布的弱监督社交网络图像标签修正和标注方法，并利用深度网络学习视觉特征和标签语义之间的关系。本论文提出的框架，不仅可以对训练图像的用户标签进行修正，同时最后的深度学习模块还可以用来对新图像进行标签标注。在两个公开数据集上的实验结果，验证了本方法的有效性，并且利用不同的尺度参数来对噪声进行拟合，初步得到了两个数据集的噪声分布情况。从结果上来说，CDNL 模型对 MIRFlickr 数据集噪声拟合非常好，表现好于主流方法，而

对 NUS-WIDE 数据集噪声拟合程度稍差于 MIRFlickr，但也取得非常不错的效果。在后续的工作中，我们可以利用更多的分布模型来更好的拟合噪声，尝试利用更为先进的网络模型，来揭示图片深层特征和标签语义的内在关系。

参考文献（References）

[1] Kennedy L S, Chang S F, Kozintsev I V. To search or to label? Predicting the performance of search-based automatic image classifiers[C]//Proceedings of the 8th ACM international workshop on Multimedia information retrieval. 2006: 249-258.

[2] Chen M, Zheng A,Weinberger K. Fast image tagging[C]//International conference on machine learning. 2013: 1274-1282.

[3] Li Z, Tang J. Weakly supervised deep matrix factorization for social image understanding[J]. IEEE Transactions on Image Processing, 2016, 26(1): 276-288.

[4] Zhu G, Yan S, Ma Y. Image tag refinement towards low-rank, content-tag prior and error sparsity[C]//Proceedings of the 18th ACM international conference on Multimedia. 2010: 461-470.

[5] Li Z, Tang J. Weakly-supervised deep nonnegative low-rank model for social image tag refinement and assignment[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[6] Li X, Snoek C G M, Worring M. Learning social tag relevance by neighbor voting[J]. IEEE Transactions on Multimedia, 2009, 11(7): 1310-1322.

[7] Tang J, Shu X, Qi G J, et al. Tri-clustered tensor completion for social-aware image tag refinement[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(8): 1662-1674.

[8] H. Ma, C. Liu, I. King, et al. Probabilistic factor models for web site recommendation[C]. Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011, 265-274

[9] Park S, Serpedin E, Qaraqe K. Gaussian assumption: The least favorable but the most useful [lecture notes][J]. IEEE Signal Processing Magazine, 2013, 30(3): 183-186.

[10] S. Park, E. Serpedin, K. A. Qaraqe. Gaussian assumption: The least favorable but the most useful[J]. IEEE Signal Processing Magazine, 2013, 30(3): 183-186

[11] Liu D, Hua X S, Wang M, et al. Image retagging[C]//Proceedings of the 18th ACM international conference on Multimedia. 2010: 491-500.

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[13] Dolan-Gavitt B, Leek T, Zhivich M, et al. Virtuoso: Narrowing the semantic gap in virtual machine introspection[C]//2011 IEEE symposium on security and privacy. IEEE, 2011: 297-312.

[14] Barnard K, Duygulu P, Forsyth D, et al. Matching words and pictures[J]. Journal of machine learning research, 2003, 3(Feb): 1107-1135.

[15] Makadia A, Pavlovic V, Kumar S. Baselines for image annotation[J]. International Journal of Computer Vision, 2010, 90(1): 88-105.

[16] Wu L, Jin R, Jain A K. Tag completion for image retrieval[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(3): 716-727.

[17] Zhao R, Grosky W I. Narrowing the semantic gap-improved text-based web document retrieval using visual features[J]. IEEE transactions on multimedia, 2002, 4(2): 189-200.

[18] Feng Z, Feng S, Jin R, et al. Image tag completion by noisy matrix recovery[C]//European Conference on Computer Vision. Springer, Cham, 2014: 424-438.

[19] Bengio Y. Learning deep architectures for AI[M]. Now Publishers Inc, 2009.

[20] Murthy V N, Maji S, Manmatha R. Automatic image annotation using deep learning representations[C]//Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. 2015: 603-606.

[21] Zhang J, Wu Q, Zhang J, et al. Kill two birds with one stone: Weakly-supervised neural network for image annotation and tag refinement[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[22] Li Z, Tang J, Zhang L, et al. Weakly-supervised Semantic Guided Hashing for Social Image Retrieval[J]. INTERNATIONAL JOURNAL OF COMPUTER VISION, 2020.

[23] Li Z, Tang J, Mei T. Deep collaborative embedding for social image understanding[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(9): 2070-2083.

[24] Chen B, Wu B, Zareian A, et al. General Partial Label Learning via Dual Bipartite Graph Autoencoder[J]. arXiv preprint arXiv:2001.01290, 2020.

[25] You R, Guo Z, Cui L, et al. Cross-Modality Attention with Semantic Graph Embedding for Multi-Label Classification[C]//AAAI. 2020: 12709-12716.

[26] Li Z, Tang J. Weakly supervised deep metric learning for community-contributed image retrieval[J]. IEEE Transactions on Multimedia, 2015, 17(11): 1989-1999.

[27] Du X, Liu Q, Li Z, et al. Cauchy Matrix Factorization for Tag-Based Social Image Retrieval[J]. IEEE Access, 2019, 7: 132302-132310.

[28] Lin Z, Chen M, Ma Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices[J]. arXiv preprint arXiv:1009.5055, 2010.

[29] Huiskes M J, Lew M S. The MIR flickr retrieval evaluation[C]//Proceedings of the 1st ACM international conference on Multimedia information retrieval. 2008: 39-43.

[30] Tang J, Shu X, Li Z, et al. Generalized deep transfer networks for knowledge propagation in heterogeneous domains[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2016, 12(4s): 1-22.

[31] Murthy V N, Maji S, Manmatha R. Automatic image annotation using deep learning representations[C]//Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. 2015: 603-606.

[32] Gong Y, Jia Y, Leung T, et al. Deep convolutional ranking for multilabel image annotation[J]. arXiv preprint arXiv:1312.4894, 2013.

[33] Verma Y, Jawahar C V. Image annotation using metric learning in semantic neighbourhoods[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012: 836-849.

[34] Trigeorgis G, Bousmalis K, Zafeiriou S, et al. A deep matrix factorization method for learning attribute representations[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(3): 417-429.