

# 基于生成式对抗网络的图像自动标注

税留成\* 刘卫忠 冯卓明

(华中科技大学 光学与电子信息学院, 湖北 武汉 430074)

(597799047@qq.com)

**摘要:** 针对基于深度学习的图像标注模型输出层神经元数目会与标注词汇量成正比,导致模型结构会因词汇量的变化而改变的问题,提出了结合生成式对抗网络(GAN)和 word2vec 的新标注模型。首先,通过 word2vec 将标注词汇映射为固定的多维词向量;其次,利用生成式对抗网络构建一个神经网络模型(GAN-W),使输出层神经元数目与多维词向量维数相等,与词汇量不再相关;最后,通过对模型多次输出的排序结果来确定最终标注。模型分别在 Corel 5K 和 IAPRTC-12 图像标注数据集上进行实验,在 Corel 5K 数据集上,模型准确率、召回率和 F1 值比卷积神经网络回归方法(CNN-R)分别提高 5 个百分点、14 个百分点、9 个百分点;在 IAPRTC-12 数据集上,模型准确率、召回率和 F1 值比两场 K 最邻近模型(2PKNN)分别提高 2 个百分点、6 个百分点、3 个百分点。实验结果表明,GAN-W 模型可以解决输出神经元数目随词汇量改变的问题,同时每幅图像标注的标签数目自适应,使得模型标注结果更加符合实际标注情形。

**关键词:** 图像自动标注;深度学习;生成式对抗网络;标注向量化;迁移学习

**中图分类号:** TP 391.41

**文献标志码:** A

## Automatic image annotation based on Generative Adversarial Network

SHUI Liu-Cheng LIU Wei-Zhong FENG Zhuo-Ming

School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan, 430074, China

**Abstract:** In order to solve the problem that the number of output neurons in deep learning-based image annotation model was directly proportionate to the labeled vocabulary, a new annotation model combining the Generative Adversarial Network (GAN) and word2vec was proposed. Firstly, the labeled vocabulary was mapped to the multidimensional word vector through word2vec; Secondly, a neural network model (GAN-W) using GAN was established and the number of neurons in the model output layer was equal to the dimensions of the word vector, no longer relevant to vocabulary; Finally, the annotation result was determined by sorting the multiple output of the model. Experiments are conducted on the image annotation datasets Corel 5K and IAPRTC-12. The experimental results show that on the Corel 5K dataset, accuracy rate, recall rate, and F1 value of the proposed model have increased by 5 percentage points, 14 percentage points and 9 percentage points respectively, compared with the Convolutional Neural Network Regression (CNN-R); On the IAPRTC-12 dataset, accuracy rate, recall rate and F1 value are 2 percentage points, 6 percentage points and 3 percentage points higher than those of the Two-Pass K-Nearest Neighbor (2PKNN). The results show that the GAN-W model can solve the issue caused by the change of neuron number in the output layer. Meanwhile, it is self-adaptive to the number of label in each image, which is more suitable for actual annotation situation.

**Keywords:** Automatic image annotation; Deep learning; Generative adversarial network; Label vectorization; Migration learning

### 0 引言

随着图像数据的快速增长,通过人工对图像进行标注已经变得不可取,迫切需要对图像内容进行自动标注,以实现对图像的有效管理与检索,更加高效利用庞大的图像信息。目前,主要的标注方法是通过机器学习构建一个图像标注模

型,通过学习图像与其对应标注之间的潜在联系,给未知图像添加描述其内容的关键词,实现对未知图像的标注。

基于机器学习的图像标注模型大致分为 3 类:生成模型、最邻近模型及判别模型。生成模型首先提取图像特征,然后计算图像特征与图像标签之间的联合概率,最后根据测试图像的特征计算各标签的概率,确定图像对应的标签。代表方法有:多贝努利相关模型(Multiple Bernoulli Relevance

收稿日期:2018-12-05; 修回日期:2019-01-21; 录用日期:2019-01-22。

**作者简介:** 税留成(1992-),男,四川成都人,硕士研究生,主要研究方向为计算机视觉、图像标注;刘卫忠(1972-),男,湖北荆州人,副教授,博士,主要研究方向为多媒体信源编码、机器学习;冯卓明(1970-),男,湖北荆州人,博士,主要研究方向为无线通信。

Model, MBRM)<sup>[1]</sup>、跨媒体相关模型(Cross Media Relevance Model, CMRM)<sup>[2]</sup>及 SKL-CRM(Sparse Kernel Learning Continuous Relevance Model)模型<sup>[3]</sup>。最邻近模型首先根据某些基于图像特征的距离找到多幅与预测图像相似的图像, 然后根据这些相似图像的标注确定预测图像的标注。代表方法有: JEC(Joint Equal Contribution)模型<sup>[4]</sup>、2PKNN(Two-Pass K-Nearest Neighbor)模型<sup>[5]</sup>、及 TagProp\_ML(Tag Propagation Metric Learning)模型<sup>[6]</sup>。

判别模型是将图像标签视作图像的一个分类, 因此图像标注可以看成是对图像的多分类, 通过图像的分类结果确定图像的标签。代表方法有: CBSA(Content-Based Soft Annotation)模型<sup>[7]</sup>、PAMIR(Passive-Aggressive Model for Image Retrieval)模型<sup>[8]</sup>、ASVM-MIL(Asymmetrical Support Vector Machine-Based MIL Algorithm)模型<sup>[9]</sup>。近几年, 随着深度学习在图像分类上取得良好效果, 深度学习的方法也逐渐应用于图像标注任务中。例如 2016 年黎健成等<sup>[10]</sup>在 CNN(Convolutional Neural Network)模型基础上增加基于 Softmax 层的多标签排名损失函数, 提出 Multi-label CNN 标注模型; 2017 年高耀东等<sup>[11]</sup>提出基于均方误差损失的 CNN-MSE(CNN-Mean Squared Error)模型; 2018 年汪鹏等<sup>[12]</sup>提出基于多标签平滑单元的 CNN-MLSU(CNN-Multi-Label Smoothing Unit)模型; 李志欣等<sup>[13]</sup>提出结合深度卷积神经网络和集成分类器链的 CNN-ECC(CNN-Ensemble Of Classifier Chains)模型。这些模型在图像标注任务上均取得了良好的效果, 性能较传统的标注方法有明显的提高。

然而, 这些深度学习标注模型有一个共同的特点, 即模型输出层神经元(或分类器)数目与标注词汇量成正比。这将导致 2 个问题: (1)随着数据集标注词汇量的增加, 输出层神经元数目会成比例的增加。当数据集词汇量较小时, 对模型几乎没有影响, 但是如果选择较大词汇量的数据集时, 模型输出层神经元数目将变得非常庞大, 如选择 Open Images 数据集神经元数目将超过 2 万。庞大的输出层神经元数目将导致很难设计出一个合理的神经网络结构, 并且会导致模型参数量的骤增, 增加模型训练难度的同时使得模型权重文件的大小骤增, 不利于模型的实际应用; (2)当标注的词汇量发生变化时, 即使只是增删某个词汇, 由于模型输出神经元数目与词汇量成正比, 所以也需要对模型网络结构进行修改。在实际应用中新增词汇几乎是不可避免的, 这将使得模型结构将会被频繁修改, 导致模型稳定性较差。

针对此问题, 本文将生成式对抗网络(Generative Adversarial Networks, GAN)<sup>[14]</sup>和自然语言处理中的 word2vec 模型相结合, 构建一种新的图像标注模型 GAN-W。模型的主要步骤是: 首先, 利用 word2vec 将标签转换为一个固定维数的多维空间向量, 多维空间向量的维数自由选择, 模型输出层神经元数目将只与多维向量的维数相关, 不再与标注词汇量相关。另外, 当词汇量发生较小变化时, 只需要修改 word2vec 的词向量转换表即可, 不再需要修改模型结

构。其次, 标注模型不再一次性输出图像对应所有标注, 而是利用 GAN 网络每次输出一个候选标注对应的多维空间向量。通过 GAN 网络中随机噪声的扰动, 使得 GAN 网络每次可以输出与图像相关并且不同的候选标注对应的多维空间向量。最终根据模型多次输出结果筛选出图像的最终标注。

## 1 生成式对抗网络

生成式对抗网络(GAN)的核心思想源于博弈论的纳什均衡<sup>[15]</sup>, 其模型如图 1 所示, 主要由一个生成器(G)和一个判别器(D)构成, 生成器通过随机噪声生成接近数据集分布的假数据, 判别器则需要辨别输入其中的数据是来源于生成器还是数据集。

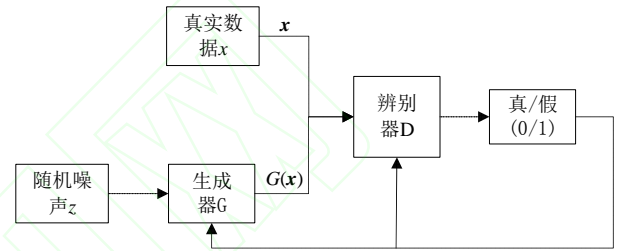


图 1 GAN 模型图

Fig. 1 GAN model

GAN 的目标函数为:

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}} [\log D(x)] + E_{z \sim P_z} [\log(1 - D(G(z)))] \quad (1)$$

GAN 网络训练时需要交替优化生成器与判别器, 优化生成器时, 最小化目标函数  $V(D, G)$ , 使生成的数据  $G(z)$  愈加接近数据集, 经过判别器后的输出  $D(G(z))$  越来越接近于 1, 即判别器无法辨别生成数据  $G(z)$  和真实数据  $x$ ; 优化判别器时, 最大化  $V(D, G)$ , 使得  $D(G(z))$  接近于 0, 同时  $D(x)$  接近于 1, 即让判别器尽可能准确判断输入数据是来自于数据集的真实数据  $x$  还是来自于生成器生成的数据  $G(z)$ 。通过多次交替优化生成器和判别器, 分别提升其性能, 最终生成器与判别器性能达到纳什均衡, 使得生成器生成的数据分布近似于原数据集的分布。

随机噪声  $z$  使得生成结果具有不确定性, 给 GAN 的生成结果带来了多样性, 与此同时, 由于缺乏约束常导致生成结果不可控。为解决这个问题, Mirza 等<sup>[16]</sup>提出条件生成对抗网络(Conditional Generative Adversarial Nets, CGAN), 在生成器输入噪声  $z$  的同时输入一个条件  $c$ , 并且将真实数据  $x$  和条件  $c$  作为判别器的输入, 利用条件  $c$  对 GAN 的生成结果进行限制。CGAN 的目标函数  $V(D, G)$ , 如式(2)所示:

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}} [\log D(x, c)] + E_{z \sim P_z} [\log(1 - D(G(z, c)))] \quad (2)$$

原始 GAN 具有训练不稳定、模式崩溃等问题,对此 Arjovsky 等<sup>[17]</sup>提出 Wasserstein-GAN (WGAN)对 GAN 进行改进,去掉判别器(D)最后 sigmoid 层,损失函数不取 log,并且对更新后的权重强制截取到一定范围。WGAN 减小了 GAN 网络的训练难度,但是 WGAN 强制截取权重容易导致模型梯度消失或者梯度爆炸。对此, Gulrajani 等<sup>[18]</sup>提出 Improved WGAN 对 WGAN 进一步改进,使用梯度惩罚代替强制截取梯度。

Improved WGAN 网络的目标函数为:

$$\min_G \max_D E_{\mathbf{x} \sim P_{data}} [D(\mathbf{x})] - E_{\mathbf{z} \sim P_z} [D(G(\mathbf{z}))] + \lambda E_{\tilde{\mathbf{x}} \sim P_{\tilde{\mathbf{x}}}} [(\|\nabla_{\tilde{\mathbf{x}}} D(\tilde{\mathbf{x}})\|_2 - 1)^2] \quad (3)$$

其中,  $E_{\tilde{\mathbf{x}} \sim P_{\tilde{\mathbf{x}}}} [(\|\nabla_{\tilde{\mathbf{x}}} D(\tilde{\mathbf{x}})\|_2 - 1)^2]$  为训练判别器 (D) 时梯度惩罚带来的损失,  $\lambda$  为梯度惩罚损失的系数,一般取值为 10。

## 2 词向量

由于神经网络无法直接处理文本数据,所以需要文本数据进行数值转换。传统的方法是将文本数据转换成 one-hot 词向量,即词向量维数与词汇量相等,所有单词均分别与向量某一维对应,并且如果单词存在,则对应维度取值为 1,否则只能为 0,如在 5 维的词向量中 cat 可能表示为[0 0 0 1 0],dog 为[0 1 0 0 0]。One-hot 表示方法是一种高维稀疏的方法,词向量维度与词汇量成正比,计算效率低而且每一维度互相正交,无法体现词之间的语义关系。

2013 年 Google 开源一款新词向量生成工具 word2vec 可以将词汇映射成为多维空间向量,如 cat 可能表示为[0.1,0.25,0.3,0.01,0.9,0.6],目前 word2vec 被大量应用于自然语言处理(NLP)任务当中。word2vec 的主要思想是具有相同或相似上下文的词汇,可能其具有相似的语义,通过学习文本语料,根据词汇上下文,将文本中的每个词汇映射到一个统一 N 维词汇空间,并使语义上相近的词汇在该空间中的位置相近,如 cat 和 kitten 对应词向量之间的空间距离小于 cat 和 iPhone 之间的距离,从而体现词汇之间的关系,从而避免 one-hot 词向量的缺点。

## 3 模型网络结构

### 3.1 模型结构

本文采用的模型结构如图 2 所示。模型整体框架采用 CGAN 网络架构,输入图像大小统一为(299,299,3),图像对应的 N 维特征向量作为条件,真实标注对应的 M 维词向量作为真实数据,根据条件和 100 维随机噪声,生成器输出 M 维向量作为生成数据。其中 CNN 特征提取模型选择 Inception-ResNetV2<sup>[19]</sup>模型,并在 ImageNet 数据集上进行预训练,去除最后分类层后采用迁移学习的方法应用到模型

中;word2vec 功能采用 genism 库的 Word2Vec 模块实现,生成的词向量维数统一为 500 维,生成器和判别器均采用全连接层,将特征向量和随机噪声/词向量分别全连接映射到不同维数后拼接,重复操作 2 次后映射到输出全连接层,输出全连接层神经元数目与词向量维数相等。本文训练 GAN 采用 Improved WGAN 模型,所以判别器输出层去除 sigmoid 激活层。

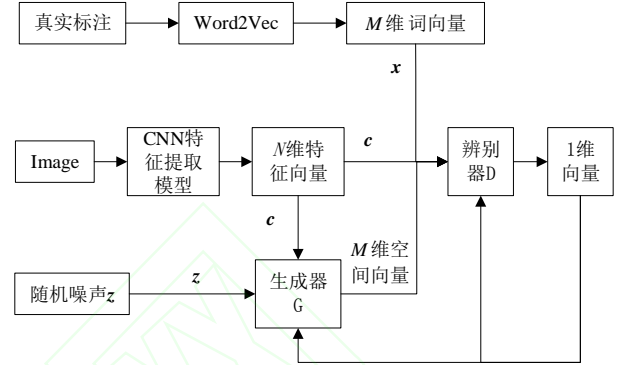


图 2 网络模型图

Fig. 2 model net

### 3.2 损失计算

在图像标注领域,标注词汇的分布不均匀是一个常见的问题,有些标注如cafe、butterfly在Corel 5K数据集中只出现过2次,而water、sky、tree等标注出现次数多于800次。由于标注中不同词汇的词频差异巨大,如果不进行处理,模型容易忽略低频标签的影响,导致对低频词汇标注的准确率下降,影响模型性能。针对标注分布不均衡问题,本模型对损失函数进行优化,对不同标注的损失乘以一个平衡系数,使得词频低的标注具有更大权重的损失,另外使用L2正则化减小模型过拟合。修改后的损失为:

$$\text{Loss}_G = -E_{\mathbf{z} \sim P_z} [D(G(\mathbf{z})) * \alpha] + \beta L \quad (4)$$

$$\text{Loss}_D = E_{\mathbf{z} \sim P_z} [D(G(\mathbf{z})) * \alpha] -$$

$$E_{\mathbf{x} \sim P_{data}} [D(\mathbf{x}) * \alpha] + \lambda E_{\tilde{\mathbf{x}} \sim P_{\tilde{\mathbf{x}}}} [(\|\nabla_{\tilde{\mathbf{x}}} D(\tilde{\mathbf{x}})\|_2 - 1)^2] \quad (5)$$

其中,  $\alpha = 10/k$ ,  $k$  为标签对应的图像数目,10 倍是防止 loss 过小,  $L$  为生成标签向量与真实标签向量的均方误差 (L2 正则化损失),  $\beta$  为  $L$  的系数。

### 3.3 标注排序

由于本文模型每次输出一个图像对应的候选标注词向量,所以本文的标注排序方法采用出现次数排序,具体过程为:1. 通过已训练模型对图像进行N次预测,获得N个词向量2. 对于每个词向量,通过Word2Vec模型获取与其对应最接近的M个候选标注词及每个标注词对应的概率3. 以标注词对应的概率作为标注词对应的出现次数,统计所有候选标注词



出现次数,通过阈值筛选出现次数大于阈值的候选标注作为该图像最终标注

## 4 实验

### 4.1 数据集

本文实验的数据集为图像标注领域常用数据集:Corel 5K 和 IAPRTC-12 数据集。Corel 5K 数据集是由科雷尔(Corel)公司收集整理的 5000 张图片,该数据集常用于图像分类、检索等科学图像实验,是图像实验的标准数据集。IAPRTC-12 数据集最初用于跨语言检索任务,每张图像有英语、德语及西班牙语三种语言的图像描述,在研究人员用自然语言处理技术提取图形描述中的常用名词作为图像标签后,也被作为图像标注任务的常用数据集。Corel 5K 和 IAPRTC-12 数据集的详细信息统计如下表:

表 1 数据集信息表  
Tab.1 Dataset information

|       | Corel 5k | IAPRTC-12  |
|-------|----------|------------|
| 图片数量  | 5000     | 19627      |
| 标签数量  | 260      | 291        |
| 测试/训练 | 500/4500 | 1962/17665 |
| 平均标签数 | 3.4      | 5.7        |

### 4.2 评估方法

实验采用的评价方法是计算数据集中每个标签的准确率( $P$ )和召回率( $R$ )及  $F1$  值。假设一个标签在测试集中相关图像为  $N$ ,测试时模型预测出的相关图像为  $N1$ ,其中预测正确的相关图像数量为  $N2$ ,那么,准确率  $P = N2 / N1$ ,召回率  $R = N2 / N$  及  $F1 = 2 * P * R / (P + R)$ 。

### 4.3 标注结果

#### 4.3.1 不同阈值对图像标注的影响

不同标注阈值对本文模型的最终标注性能有巨大影响,为了进一步探究不同阈值与标注性能的关系,本文对不同阈值下的模型的标注性能进行测试。图 3 及图 4 为模型标注的准确率、召回率、 $F1$  值与阈值的关系图。测试时,模型预测次数为 128 次,每次选出最接近输出向量的 5 个候选标,统计所有候选标注,选出出现次数大于阈值的标注作为图像最终标注。

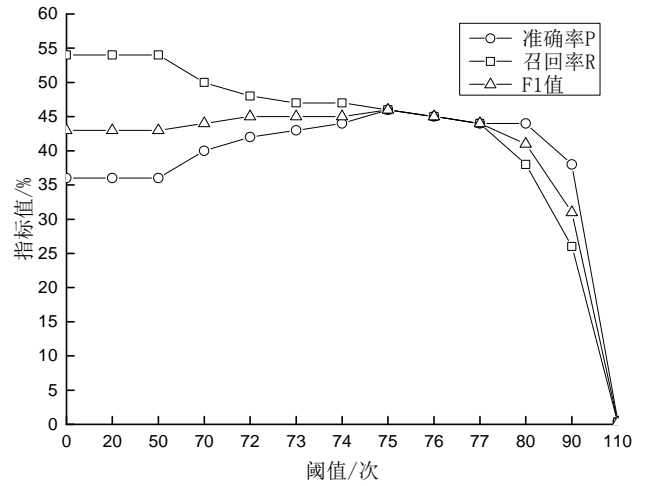


图3 Corel 5K数据集下阈值影响

Fig.3 Threshold effect under Corel 5K

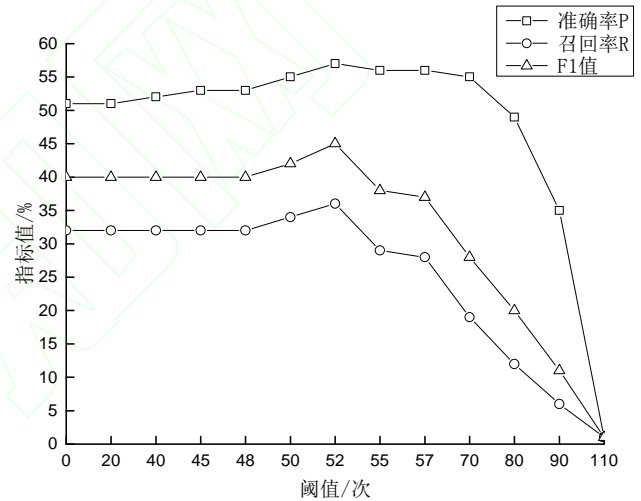


图4 IAPRTC-12数据集下阈值影响

Fig.4 Threshold effect under IAPRTC-12

从图 3 和图 4 可以看出:标注的准确率  $P$  随阈值先上升后下降,召回率  $R$  随阈值上升而下降, $F1$  值基本上随阈值略微上涨后下降。出现这种现象的原因:模型可以学到图像特征与标签向量之间的映射关系,通过对模型的训练,模型有了一定的标注能力,对于大多数标签的预测结果中,正确的预测对应的出现次数一般较高。当阈值特别小时,标签对应的出现一般次数大于阈值,标签的预测结果基本没有被阈值过滤,标注准确率  $P$  和召回率  $R$  都不变;阈值增加到一定值时,部分错误的预测被逐渐过滤,正确的预测因为出现次数较大,基本不受影响,准确率  $P$  上升,召回率  $R$  基本不变。阈值继续增加,正确的预测也开始被过滤,但是由于正确的预测情形多集中于出现次数较高的情形,因此阈值的增加对正确的预测影响更大,正确预测的部分被过滤的速度大于错误预测的部分,最终使得标注准确率  $P$  和召回率  $R$  都减小,直到正确的预测被阈值完全过滤掉,标注准确率  $P$  和召回率  $R$  都为 0。 $F1$  值的变化由准确率  $P$  和召回率  $R$  的变化共同确定。模型性能随阈值变

化,为了和其它模型标注性能进行对比及模型实际标注效果展示,需要确定模型的最佳阈值。由于  $F1$  值能兼顾准确率  $P$  和召回率  $R$ ,所以  $F1$  值作为模型最佳阈值选取的参考,选取  $F1$  值最大时的阈值作为模型最佳阈值。由于不同数据集之间存在差异导致对于不同数据集模型的最佳阈值也不相同,所以对于 Corel 5K 和 IAPRTC-12 数据集,在模型预测次数为 128 次的情况下,模型分别选择 75 和 50 作为模型的最佳阈值。

#### 4.3.2 不同模型标注性能对比

本文将 GAN-W 模型与其他经典的标注方法进行对比,来验证本文所提出模型的有效性。这里涉及的方法包括:传统模型方法 RF-opt(Random Forest-Optimize)<sup>[20]</sup>、2PKNN<sup>[5]</sup>、2PKNN-ML(2PKNN-Metric Learning)<sup>[5]</sup>、SKL-CRM<sup>[3]</sup>、KSVM-VT<sup>[21]</sup> 和使用深度卷积神经网络的方法 NN-CNN(NearestNeighbor-CNN)<sup>[22]</sup>、CNN-R (CNN-Regression)<sup>[23]</sup>、ADA(Attribute Discrimination Annotation)<sup>[24]</sup>、SNDF(Automatic Image Annotation Combining Semantic Neighbors And Deep Features)<sup>[25]</sup>、CNN-MSE<sup>[11]</sup>、CNN-MLSU<sup>[12]</sup>。表 2 显示本文 GAN-W 模型与其它模型在 Corel 5K 和 IAPRTC-12 数据集上标注性能的对比。

表 2 模型性能对比表

| Tab.2 Comparison of model performance |           |           |           |           |           |           |
|---------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|
| 模型                                    | Corel 5K  |           |           | IAPRTC-12 |           |           |
|                                       | R         | P         | F1        | R         | P         | F1        |
| RF-opt                                | 40        | 29        | 34        | 31        | 44        | 36        |
| 2PKNN                                 | 40        | 39        | 39        | 32        | 49        | 39        |
| 2PKNN-ML                              | 46        | 41        | 43        | 32        | 53        | 40        |
| SKL-CRM                               | 46        | 39        | 42        | 32        | 51        | 39        |
| KSVM-VT                               | 42        | 32        | 44        | 29        | 47        | 36        |
| NN-CNN                                | 45        | 42        | 44        | 32        | 54        | 41        |
| CNN-R                                 | 41        | 32        | 37        | 31        | 49        | 38        |
| ADA                                   | 40        | 32        | 36        | 30        | 42        | 35        |
| SNDF                                  | 39        | 37        | 38        | 30        | 48        | 37        |
| CNN-MSE                               | 35        | 41        | 38        | 35        | 40        | 37        |
| CNN-MLSU                              | <b>49</b> | 37        | 42        | <b>38</b> | 44        | 41        |
| GAN-W                                 | 46        | <b>46</b> | <b>46</b> | 34        | <b>55</b> | <b>42</b> |

通过表 2 可以看出,本文提出的 GAN-W 模型在 Corel 5K 数据集上,性能较传统方法有了较大提高,召回率取得并列第一,高于 RF-opt 方法 4%,准确率和  $F1$  值均为第一,比 RF-opt 方法分别提高 17% 和 12%,在使用卷积模型的方法中,召回率比 CNN-MSE 方法提高了 11%,取得第二高的召回率,准确率和  $F1$  值均为第一。在 IAPRTC-12 数据集上,模型也有良好表现,准确率和  $F1$  值均为第一,召回率也取得不错效果。综合 GAN-W 模型在 Corel 5K 和 IAPRTC-12 数据集上的性能指标数据可以得出,GAN-W 模型与其它的方法相比,虽然召回率低于 CNN-MLSU 方法未取得最高值,但是效果

依然良好,同时模型准确率和  $F1$  值均取得较大提升,取得最佳效果,模型的综合性能与其他模型相比具有明显的提高。

#### 4.3.3 模型实际标注效果

图 5 中给出模型自动标注的实际结果,模型统一预测次数为一个 batch\_size,128 次,测试 Corel 5K 数据集时选择的阈值为 75,每幅图像选取出出现次数大于阈值的标注作为该图像最终标注。






| 图像   | 原始标注                         | 模型标注                         |
|--|------------------------------|------------------------------|
|    | city,sun,water               | city,sun,water               |
|    | jet,plan,f-16                | jet,plan,f-16                |
|   | coral, anemone, ocean, reefs | coral, anemone, ocean, reefs |
|  | tree, horses, mare, foals    | tree, horses, mare, foals    |
|  | bear, polar, snow, face      | bear, polar, snow, tundra    |

图 5 Corel 5K 数据集下模型实际标注效果

Fig.5 Actual annotation effect of model

从图中可以看出:(1)与大部分标注模型固定每幅图像的标注数目不同,本文模型对每幅图像的标注数目不是定值,不同图像可能有不同的标注数目,更符合实际标注情况。通过对 GAN-W 模型的训练,模型可以学到图像特征与标签向量之间的映射关系,在每次预测新图像时,模型就会根据被预测图像的视觉特征中的某种特征输出一个与之对应的标签向量。对于语义简单的图像,其图像视觉特征只包含某个的标签对应的特征,所以模型每次输出的向量基本上都接近该标签,使得该标签对应的出现次数较高,而其它标签出现次数小于阈值被过滤掉,模型最终标注数目较少;对于复杂的图像,其图像视觉特征可能包含多个标签对应的特征,经过随机噪声的扰动,使得多个标签中每个标签都有较大概率成为模型输出标签,所以通过多次测试之后,多个标签中的每个标签出现次数都不会太小,模型最终的标注数目较多。(2)某些标注虽然与原标注不符合,但是可能与测试图像的语义

相符或者相关,这是因为某些标注之间(如 tundra 与 bear、snow、polar)在数据集中共现频率较高,使得这些标注在使用 Word2vector 进行向量化时,他们对应的多维向量之间的距离很近,所以在获取输出向量对应最接近的标注词时常一起出现,并且标注词之间对应的概率相差很小,导致某些标注虽然不是原始标注,但是最终统计次数时出现次数依然很大,被确定为图像标注之一。同时,由于在数据集中这些标注经常一起出现,证明在现实中它们之间的联系较深,所以在新的测试图像中,这些常与原始标注一起出现的标签依然有较大概率与测试图像相关。例如上表中的 tundra 不在原始标注中,但是 tundra 在数据集中多与 bear、snow、polar 一起出现,所以 tundra 被作为最终输出之一,依然与图像内容有联系。

## 5 结语

针对基于深度学习的图像自动标注模型其结构受标注词汇量影响的问题,本文基于生成式对抗网络和词向量模型提出一种新标注模型 GAN-W,通过在 Corel 5K 和 IAPRTC-12 数据集上的实验结果表明 GAN-W 模型的准确率  $P$ 、召回率  $R$  及  $F1$  值较其他模型有明显的提高,证明本文模型能够较好的应用于图像标注任务,标注结果更加符合实际标注情况。然而,模型存在一些值得改进和研究的方面:(1)词向量的训练结果缺乏一个较好的评判标准(2)生成器和判别器的网络模型需要进行进一步优化(3)选择更优的特征提取模型和标签平衡系数。

## 参考文献

- [1] Feng S L, Manmatha R, Lavrenko V. Multiple Bernoulli relevance models for image and video annotation[C]// Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision & Pattern Recognition. Washington, D.C.: IEEE, 2004:1002-1009.2004.
- [2] JEON J, LAVRENKO V, MANMATHA R. Automatic image annotation and retrieval using cross-media relevance models[C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM, 2003:119-126
- [3] MORAN S, LAVRENKO V. A sparse kernel relevance model for automatic image annotation[J].Journal of Multimedia Information Retrieval,2014, 3(4):209-229
- [4] MAKADIA A, PAVLOVIC V, KUMAR S. Baselines for image annotation[J]. International Journal of Computer Vision, 2010, 90(1):88-105
- [5] VERMA Y, JAWAHAR C V. Image annotation using metric learning in semantic neighborhoods[C]//Proceedings of the 12th European Conference on Computer Vision. Berlin:Springer, 2012:836-849.
- [6] GUILLAUMIN M, MENSINK T, VERBEEK J, et al. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation[C]//Proceedings of the 12th IEEE International Conference on Computer Vision. Piscataway, NJ:IEEE, 2009:309-316.
- [7] Chang E, Goh K, Sychay G, et al. CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines [J]. IEEE Transactions on Circuits & Systems for Video Technology, 2003, 13(1):26-38.
- [8] Grangier D, Bengio S. A Discriminative Kernel-Based Approach to Rank Images from Text Queries[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2008, 30(8):1371-1384.
- [9] Yang C, Dong M, Hua J. Region-based Image Annotation using Asymmetrical Support Vector Machine-based Multiple-Instance Learning[C]// Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision & Pattern Recognition. Washington, D.C.:IEEE,2006:2057-2063.
- [10] 黎健成,袁春,宋友.基于卷积神经网络的多标签图像自动标注[J].计算机科学, 2016, 43(7):41-45.(LI J C, YUAN C, SONG Y. Multi-label image annotation based on convolutional neural network[J]. Computer Science, 2016, 43(7):41-45.)
- [11] 高耀东,侯凌燕,杨大利.基于多标签学习的卷积神经网络的图像标注方法[J].计算机应用, 2017, 37(1):228-232.(GAO Y D, HOU L Y, YANG D L. Automatic image annotation method using multi-label learning convolutional neural network[J].Journal of Computer Applications,2017, 37(1):228-232.)
- [12] 汪鹏, 张奥帆, 王利琴, 董永峰. 基于迁移学习与多标签平滑策略的图像自动标注[J]. 计算机应用, 2018, 38(11): 3199-3203. (WANG P, ZHANG A F, WANG L Q, DONG Y F. Image automatic annotation based on transfer learning and multi-label smoothing strategy. Journal of Computer Applications, 2018, 38(11): 3199-3203.)
- [13] 李志欣, 郑永哲, 张灿龙, et al. 结合深度特征与多标记分类的图像语义标注[J].计算机辅助设计与图形学学报, 2018, 30(2): 318-326. (LI Z X, ZHENG Y Z, ZHANG C L, et al. Combining Deep Feature and Multi-label Classification for Semantic Image Annotation. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(2): 318-326.)
- [14] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proceedings of the 2014 Conference on Advances in Neural Information Processing Systems 27. Montreal, Canada: Curran Associates, Inc., 2014. 2672-2680
- [15] 王坤峰, 苟超, 段艳杰, et al. 生成式对抗网络 GAN 的研究进展与展望[J]. 自动化学报, 2017, 43(3):321-332. (WANG K F, GOU C, DUAN Y J, et al. Generative adversarial networks: the state of the art and beyond[J]. ACTA Automatica Sinica, 2017, 43(3): 321-332.)
- [16] Mirza M, Osindero S. Conditional Generative Adversarial Nets[J]. arXiv preprint arXiv:1411.1784,2014.
- [17] Arjovsky M, Chintala S, Bottou, Léon. Wasserstein GAN[J].arXiv preprint arXiv:1701.07875, 2017.
- [18] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved Training of Wasserstein GANs[C]. Proceedings of the 30th Advances in Neural Information Processing Systems, California: NIPS,2017:5769-5779.
- [19] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[C]. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco: AAAI, 2017:4278-4284.
- [20] Fu H, Zhang Q, Qiu G. Random forest for image annotation[C]// Proceedings of the 12th European conference on computer vision, Berlin: Springer, 2012:86-99
- [21] Verma Y, Jawahar C. Exploring SVM for image annotation in presence of confusing labels[C]// Proceedings of the 24th British machine vision conference, London : BMVA Press,2013: 1-11.
- [22] KASHANI M M, AMIRI S H. Leveraging deep learning representation for search-based image annotation[C]//Proceedings of 2017 Artificial Intelligence and Signal Processing Conference. Piscataway, NJ:IEEE, 2017:156-161.
- [23] 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. New York:ACM, 2015:603-606.

- [24] 周铭柯, 柯道, 杜明智. 基于数据均衡的增进式深度自动图像标注. 软件学报, 2017, 28(7): 1862-1880. (ZHOU MK, KE X, DU MZ. Enhanced deep automatic image annotation based on data equalization. Journal of Software, 2017, 28(7): 1862-1880)
- [25] 柯道, 周铭柯, 牛玉贞. 融合深度特征和语义邻域的自动图像标注[J]. 模式识别与人工智能, 2017, 30(3):193-203. (KE X, ZHOU M K, NIU Y Z. Automatic Image Annotation Combining Semantic Neighbors and Deep Features. Pattern Recognition and Artificial Intelligence, 2017, 30(3): 193-203.)

SHUI Liucheng, born in 1992, M. S. candidate. His research interests include deep learning、image annotation.

LIU Weizhong, born in 1972, Ph. D, associate professor. His research interests include Multimedia source coding and machine learning.

FENG Zhuoming, born in 1970, Ph. D, His research interests include wireless communication