

分类号_____

学校代码 10487

学号 M201672055

密级_____

华中科技大学

硕士学位论文

基于生成式对抗网络的图像标注方法 研究

学位申请人： 税留成

学 科 专 业： 微电子学与固体电子学

指 导 教 师： 刘卫忠 副教授

答 辩 日 期： 2019 年 5 月 16 日

**A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of**

**Research on Image Annotation Based on Generative
Adversarial Network**

Candidate : Shui Liucheng

**Major : Microelectronics and
Solid-state Electronics**

Supervisor : Prof. Liu Weizhong

Huazhong University of Science & Technology

Wuhan 430074, P.R.China

May, 2019

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 ☐， 在 _____ 年解密后适用本授权书。
本论文属于 不保密 ☐。

（请在以上方框内打“√”）

学位论文作者签名：

日期： 年 月 日

指导教师签名：

日期： 年 月 日

摘要

随着互联网技术的发展以及智能移动设备的快速普及,每天都会产生数以亿计的图像数据并且被各个用户上传到互联网,这些图像数据在大多杂乱无序的同时又包含着海量有用信息。为了对这些图像数据进行有效管理并高效利用其包含的有用信息,图像语义自动标注技术应运而生。

目前,图像自动标注技术大多通过传统机器学习或者深度学习的方法构建标注模型实现对未知图像的自动标注。但是,这些标注方法大多都存在一个问题,即输出层的神经元(分类器)数目与数据集标注词汇量成比例,这将导致2个问题:1.模型实用性较差,当数据集词汇量较大时,过大的输出层数目将会急剧增加模型的设计和训练难度;2.模型结构稳定性差,模型结构会随词汇量变化而改变。

针对上述问题,本文将生成式对抗网络与 Word2vec 词向量模型相结合,设计并实现了一种新的标注模型。首先,通过 Word2vec 模型将标注词汇映射为一个维数固定且可选择的多维词向量;其次,利用生成式对抗网络构建一个神经网络模型(GAN-W),使模型生成器的输出层神经元数目与多维词向量维数相等,生成器将生成与词向量同维度的向量,使模型输出层神经元数目与标注词汇量解绑;最后,对模型多次输出结果进行排序,通过排序结果来决定图像对应的最终标注。

本文模型分别在 Corel 5K 和 IAPRTC-12 图像标注数据集上进行了实验:1.通过 Word2vec 模型输出向量维度对模型性能影响的实验证明了本文模型能够解决上述问题,模型的输出神经元数目可以在一个很大范围内自由选择。2.通过与其他模型的性能对比实验得出本文模型的准确率 P 和 F1 值均高于其它模型,同时召回率 R 仅次于 CNN-MLSU 模型,模型的标注性能具有较大的提升。3.通过模型的实际标注结果展示出了本文模型对每幅图像标注的标签数目自适应,更加符合实际标注情况。总而言之,本文模型在解决输出层神经元数目与数据集标注词汇量成比例问题的同时模型相较于其它标注模型标注在标注性能上有一定提高,在实际标注结果中同样具有优势。

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

Abstract

With the development of Internet technology and the rapid popularization of smart devices, the hundreds millions of image data were generated every day and uploaded to the Internet, most of these image data were cluttered and contained a lot of information. In order to manage these image data and use the information, the image automatic annotation technology was proposed.

At present, most of the image automatic annotation technology build the annotation model through traditional machine learning or deep learning methods to generate labels of unknown image. However, most of these annotation methods have a problem that the number of neurons (classifiers) in the output layer is directly proportionate to the label vocabulary, which will lead to two problems: 1. When the label vocabulary is very large, the annotation model will be less practical. Because the huge output layer will increase the difficult of the model design and training; 2. The structure stability of the annotation model is poor, the model structure will change with the label vocabulary.

In order to solve the problem, a new annotation model combining the Generative Adversarial Network (GAN) and Word2vec is designed and implemented. First, the label is mapped to a fixed and optional multidimensional word vector by the Word2vec model. Secondly, a neural network model (GAN-W) is built using the GAN and the neurons number of the model output layer was equal to the dimensions of the word vector. So, the model generator will generate a vector with the same dimension of the word vector, no longer relate to the label vocabulary. Finally, by sorting the output of the model, the final label of image is determined .

Experiments are conducted on the Corel 5K and IAPR TC-12 image annotation data set, the results show that: 1. The experiment of the vector dimension influence on the annotation performance proves that the model can solve the above-mentioned problem and the neurons number of the output layer can be freely selected in a wide range. 2. Comparing with other classical model performance , it is shown the accuracy R and F1 values are higher than other classical model, at the same time, the recall rate R is second only to the

CNN-MLSU model, the annotation performance of the model has a large improvement. 3.The actual label results of the model show it is self-adaptive to the number of label in each image, which is more suitable for actual annotation situation. In summary, the model proposed in this paper can solve the problem that the neurons number of the output layer is directly proportionate to the label vocabulary, meanwhile the model performance has a improvement compared with other classical model, and it also has advantages in the actual annotation situation..

Key words: Automatic image annotation ; Deep learning ; Generative adversarial network; Label vectorization; Transfer learning

目录

1 绪论.....	1
1.1 论文研究背景和意义.....	1
1.2 国内外研究现状与发展.....	2
1.3 论文的研究内容及组织机构.....	4
2 关键基础技术研究.....	6
2.1 深度学习与卷积神经网络.....	6
2.2 生成式对抗网络.....	8
2.3 词向量模型.....	11
2.4 迁移学习.....	13
2.5 本章小结.....	14
3 基于 GAN 的标注模型结构.....	15
3.1 标注模型设计思想.....	15
3.2 标注模型整体结构.....	16
3.3 数据预处理.....	17
3.4 GAN 网络训练模型.....	20
3.5 模型损失.....	23
3.6 测试模块.....	24
3.7 GAN 模型训练算法.....	25
3.8 本章小结.....	27
4 模型实验及结果分析.....	28

4.1 实验数据集.....	28
4.2 评估方法.....	28
4.3 Word2vec 模型训练参数对模型标注性能的影响.....	29
4.4 词向量维数对模型性能的影响.....	30
4.5 不同阈值对图像标注的影响.....	33
4.6 不同模型标注性能对比.....	36
4.7 模型实际标注效果.....	38
4.8 本章小结.....	40
5 总结与展望.....	42
致谢.....	44
参考文献.....	45
攻读硕士学位期间发表学术论文情况.....	50

1 绪论

1.1 论文研究背景和意义

随着数字多媒体技术日益提高、互联网用户的快速增加以及相机、移动手机等设备的广泛应用,每天都会新产生数以亿计的图像数据,这些图像数据大多被上传到互联网。截止到 2013 年 9 月份,在 Facebook 社交网站上,用户每天平均上传 3.5 亿多张图片,上传的图像总量已经超过 2500 亿张^[1]。在 2016 年的 Code 大会上,凯鹏华盈(KBCP)的互联网趋势报告显示:通过 Snapchat、Instagram、Facebook 等平台,2015 年每天上传到网络上的图片数量接近 33 亿张,而在 2013 年和 2014 年,该数据分别是 13 亿张和 20 亿张,增长速度也同样令人吃惊^{[2][3]}。这些图像数据数量庞大,增长迅速,在由社交网站用户上传至互联网上时大多数都没有对其内容进行说明,也就意味着这些数据大多杂乱无章。与此同时,这些图像数据又覆盖了我们的日常生活,其中包含的各种信息具有很大的潜在利用价值。面对如此多并且还在快速增加的图像数据,通过人工方法对其内容进行标注说明,在时间上和资源成本上都变得不可能,因此研究如何高效的对图像的内容进行自动标注是一件十分必要而且有意义的工作。

当前,在对图像进行自动标注的方法中,通过机器学习构建标注模型的方法是主要的实现方式,其主要流程是通过机器学习的方法构建一个标注模型,通过数据集学习标注图像与标注词之间的潜在联系,然后利用学习到的模型给未知图像添加可以描述其内容的关键词。基于机器学习的图像自动标注方法经过研究人员的多年研究,已经提出了许多经典的算法模型,如校准标记排序算法(CLR)^[4]、多标记分类问题的核方法(RankSvm)^[5]、多标记懒惰学习方法(MLkNN)^[6]等,这些标注模型在图像标注数据集上表现出了不错的性能。

在最近几年,随着数据集数据量的增大和计算机软/硬件性能的提升,机器学习的一个分支--深度学习成为了一个热门研究方法。深度学习的模型具有数据特征提取简单、模型适应性强、模型简单等诸多优点,使得深度学习被广泛应用于图像、语音、文本等多个领域,并且在各个领域都取得优异性能。深度学习在图像领域的良好表现使得其也开始被引入到图像标注工作之中,最初只是为传统机器

学习的标注方法提供一个自动提取图像特征的模块，后来也逐渐出现了许多全部由深度学习算法实现的图像标注模型，如 Multi-label CNN 模型^[7]、CNN-MSE 模型^[8]、CNN-ECC 模型^[9]等。这些基于深度学习的标注模型不仅取得和基于传统机器学习方法的标注模型相同、甚至更好的标注性能，同时还保持了深度学习方法的优点，逐渐开始成为图像标注领域的一个新的研究方向。

但是，通过对这些标注模型结构的仔细分析发现在这些基于深度学习的标注模型和部分基于传统机器学习的标注模型中，大部分都存在一个共同的问题，即标注模型的输出层神经元或者分类器的数量和图像标注数据集中的标注词汇量成比例。由于平时实验研究用的数据集都较小而且数据集基本不会发生改变，所以我们通常忽略这个问题对我们标注模型的影响。但是，由于这些模型的输出层的神经元（分类器）数目与数据集标注词汇量成比例，这将会导致 2 个问题：（1）当数据集比较大时，模型输出层神经元数目将随之变得非常庞大，如选择 Google 的 Open Images 数据集模型输出层的神经元数目将超过 2 万。庞大的输出层神经元数目会导致诸多问题，如：设计出一个结构合理的神经网络难度将会增加；模型参数量也会随之骤增，训练时间和训练难度都会提高；标注模型更易产生过拟合；模型权重文件的大小剧增，不利于模型在各种平台上的推广应用；（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，也会导致原来的模型结构失效，需要对模型网络结构进行修改，重新训练标注模型。而在实际应用中，由于各种原因对数据集进行增删几乎是不可避免的，这将使得模型结构将会被频繁修改，需要花费大量时间重新训练新模型，导致模型的适用性较差。所以，研究如何在利用深度学习优点进行图像标注工作的同时避免模型输出层的神经元（分类器）数目与数据集标注词汇量成比例的问题出现是一件十分有意义并且必要的工作。

1.2 国内外研究现状与发展

目前，图像自动标注领域大多采用基于机器学习的方法构建标注模型来实现对未知图像的自动标注工作，许多经典的方法、模型也被研究人员陆续提出。总体上，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。

生成模型的标注方法大致流程是首先对数据集中的图像提取图像特征，基于机器学习方法建立标注模型，然后模型通过数据集学习到图像特征与对应标签之间的联合概率，最后学习好的模型根据未知图像的图像特征计算词汇集中各标签与图像对应的概率，以此来确定未知图像对应的标签。生成模型的代表方法有：Jeon 等人提出的跨媒体相关模型(Cross Media Relevance Model, CMRM)^[10]采用分割图像区域来表示图像，通过学习关键词和聚类区域的联合概率分布为图像标注若干关键词^[11]；Lavrenko 等人对 CMRM 模型进行进一步改进得到连续空间相关模型(Continuous-space Relevance Model, CRM)^[12]，CRM 模型能对连续的特征建立学习模型并且不依靠对特征向量进行聚类；Feng 等人在 CRM 模型的基础上提出了多贝努里相关模型(Multiple Bernoulli Relevance Model, MBRM)^[13]，该模型使用多贝努里分布代替 CRM 中的多项分布来估计关键词概率，使用无参核密度函数估计图像区域特征的概率，获得了较好的标注效果^[11]。最近，Moran 等人提出一种改进的连续相关模型 SKL-CRM(Sparse Kernel Learning Continuous Relevance Model)^[14]，通过学习特征核之间的最优组合，提升图像标注性能。

最邻近模型的标注方法首先根据某些基于图像特征的图像距离计算方式，找到多幅与测试图像距离最近的相似图像，然后根据这些相似图像的标注来确定预测图像的标注。代表方法有：Makadia 等人提出 JEC(joint equal contribution)模型^[15]，将标注问题视为检索问题，JEC 利用全局低层图像特征和基本距离度量的简单平均寻找预测图像的最近邻，然后使用一种贪心的标签传递机制将关键词赋予预测的图像，取得了很好的标注精度和检索性能^[16]。TagProp(Tag Propagation)模型^[17]也是最邻近模型的一种经典代表方法，它通过将最近邻居的标签存在和不存在情况与权重进行结合，改善了标注性能；2PKNN_ML^[18]方法则是在找到测试图片的最近邻居后通过度量学习的方法优化特征间距离权重取得了良好的标注效果。Zhang 等^[19]在 k 近邻算法基础上提出了多标签分类算法 ML-kNN，首先确定测试样本的 k 个近邻样本，再利用最大后验概率来确定测试样本的标签。

判别模型是将图像对应的标签看成是对图像的一个分类类别，所以对图像进行标注就可以视作是对图像进行多分类，通过模型对每个标签训练一个对应的分类器，这样通过所有分类器的分类结果就可以确定图像对应的标签。判别模型的

代表方法有：GAO^[20]等提出对于每个标注将数据集中与该标注相关的图像和不相关的图像作为正负样本，通过 SVM 来训练一个关于该标注的分类器，将图像标注转换成基于 SVM 的一个多分类问题。chang 等人提出 CBSA(Content-Based Soft Annotation)算法^[21]，该算法基于 SVM 和 BPM(Bayes point machine)构建贝叶斯点分类机作为模型的分类器，通过训练与标签数目相同的分类器对图像进行标注。Carneiro^[22]等提出 SML 算法，利用 MIL (Multiple Instance Learning) 算法建立每个标签对应的模型，将图像标注转换为对图像的多分类，通过用贝叶斯决策规则选择最小错误率的标注为图像的最终标注。

近几年，深度学习在语音、图像、文字等多个领域都取得了良好的性能表现，不仅在 ImageNet 等比赛上取得成功，也开发了许多实际应用，如：Facebook 利用深度学习方法开发了一个让盲人可以“看到”图片内容的移动应用；微软开发的翻译软件 Skype，可以实时对多种语言的语音进行翻译。深度学习在各个领域取得的巨大成功，使得研究人员在图像标注模型的实现中开始抛弃传统机器学习方法，完全采用深度学习方法构建标注模型，例如 2015 年 Alexis Vallet 等^[23]在改进多标签损失和 CNN 模型预测方法的基础上提出了一种新标注模型，其中测试模型全由 CNN 网络实现，并 Pascal VOC 2007 数据集上取得了良好的性能；2016 年黎健成等^[8]在 CNN 模型基础上新设计了一种基于 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^[8]对卷积神经网络的损失函数进行改进，提出基于均方误差损失的 CNN-MSE 模型；2018 年汪鹏等^[24]在多标签平滑单元的基础上提出 CNN-MLSU 标注模型；李志欣等^[9]基于 CNN 网络和集成分类器链提出图像自动标注模型 CNN-ECC。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，但是对这些基于深度学习的自动标注模型和部分基于传统机器学习的标注模型都存在输出层的神经元或者分类器数目都与数据集标注词汇量成比例的问题，这个问题值得我们去研究。

1.3 论文的研究内容及组织机构

本文主要内容为构建一种新的图像标注模型，在利用深度学习优点的同时解决模型输出层的神经元（分类器）数目与数据集标注词汇量成比例的问题，并在图像标注数据集上对模型的标注性能进行多方面的验证和分析。

本文组织结构安排如下：

第一章主要对图像标注的研究背景和当前国内外研究进展进行简单介绍，对当前标注模型结构进行分析提出本文主要解决的问题及其意义，具体为：1.从近年图像数据爆炸增长的趋势和基于深度学习的标注模型缺陷的分析阐述了本文要解决的问题和意义 2.对当前图像标注技术的发展现状进行总结介绍 3.对本文的组织结构进行总结。

第二章主要对本文模型所使用的关键技术进行介绍，具体内容为：1.对深度学习及卷积神经网络进行简单介绍 2、详细阐述生成式对抗网络的原理及其缺点，引入对 CGAN 和 WAGN 模型的介绍。3.从 one-hot 向量引入对 Word2vec 词向量模型的介绍 4.对模型特征提取模块使用的迁移学习进行介绍

第三章主要对本文所设计的标注模型进行介绍，具体内容为：1、对本文模型的整体结构进行介绍 2、对模型的数据处理、训练模型结构、损失函数设置等进行详细介绍 3、给出本文模型的训练流程算法和测试流程算法

第四章主要对模型的实际性能进行实验，验证本文模型的有效性，主要内容为：1、对实验数据集及性能评价指标进行介绍 2、通过 Word2vec 参数对模型性能影响的实验，证明本文模型输出层神经元数目可以在大范围内自由选取 3、通过实验分析阈值对模型性能的影响及其原理，给出阈值参数的设置参考方式 4、通过与其他模型的标注性能进行对比，验证本文模型的有效性 5、通过具体的标注结果展示，说明本文模型在实际标注情况中的优势

第五章主要内容为对本文的研究内容及研究结果进行概括性总结并且对本文模型提出一些未来的进一步改进和研究的方向

2 系统关键技术研究

2.1 深度学习与卷积神经网络

深度学习是机器学习领域最近几年的一个新兴热门的建模方法，其概念最早由多伦多大学的 G.E.Hinton 等在 2006 年提出，指基于样本数据通过一定的训练方法得到包含多个层级的深度网络结构的机器学习过程^{[25][26]}。与浅层机器学习方法相比，深度学习的“深度”主要是体现在神经网络的层数上，浅层机器学习方法如支持向量机 SVM^[27]和集成学习 boosting 算法^[28]可以看成只有一个隐含层，逻辑回归方法（Logistic Regression, LR）^[29]则没有隐含层，经典的 BP 神经网络也只有一个隐含层，最多几个隐含层，BP 神经网络模型如图 2-1。

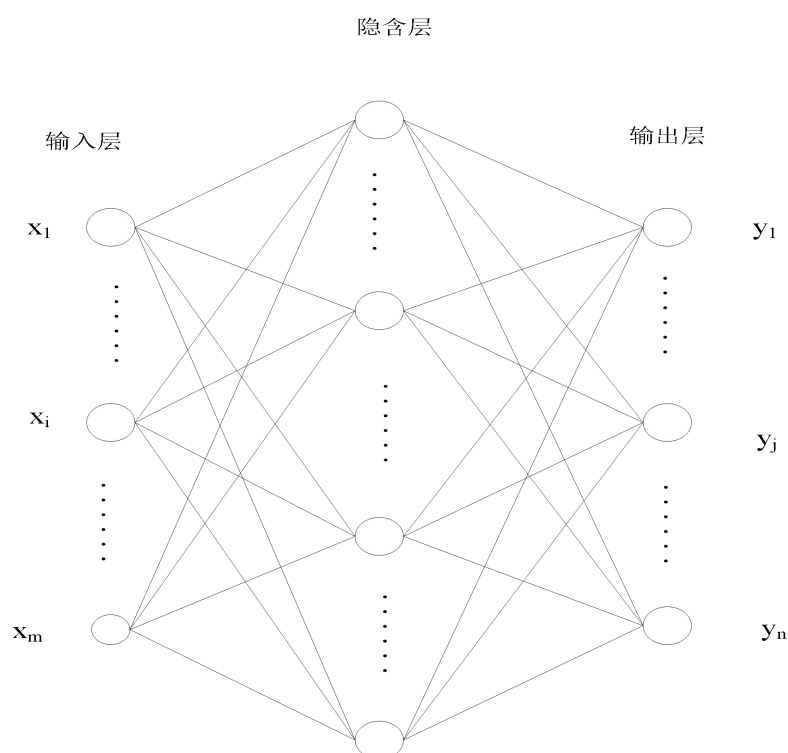


图 2-1 经典 BP 网络

浅层机器学习方法一般是只包含少数几层隐含层的浅层网络，而深度学习模型的网络层数要多很多，以下图 2-2 中 VGG16 模型为例，VGG16 为由牛津大学在 2014 年提出的经典深度学习模型，在 2014 年的 ILSVRC 比赛取得优异表现，其隐含层由 13 个卷积层和 3 个全连接层构成。深度学习模型的隐含层数基本都在 10

层以上，多的甚至达到数百层，更多的隐含层数使得深度学习的模型能更容易反映出输入和输出之间的复杂关系。

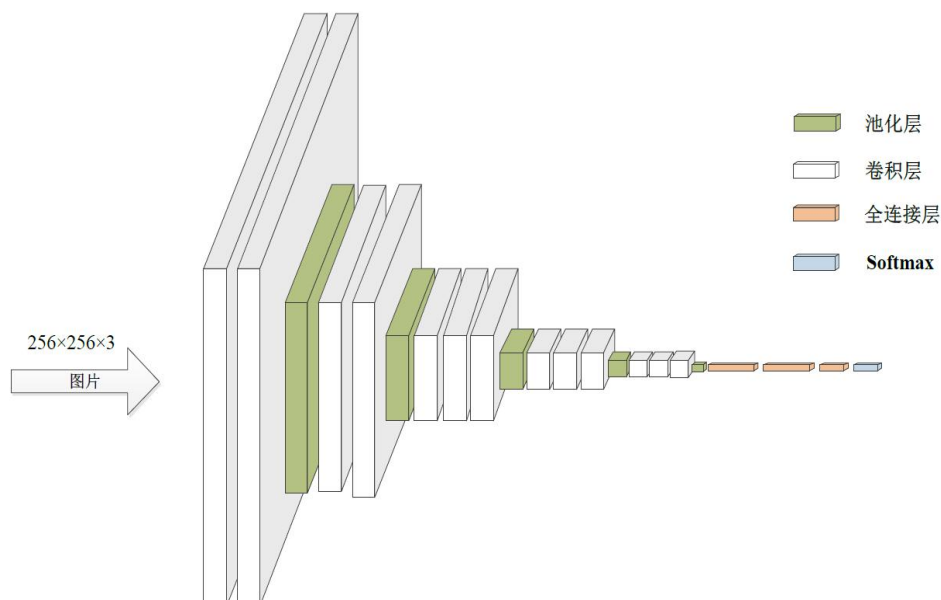


图 2-2 VGG16 网络

其实在 Hinton 提出深度学习之前，由多个隐含层构成的多隐层神经网络（MNN）就已经出现，但是由于当时采用 BP 反向传播算法和梯度下降的优化策略的问题，导致网络常出现梯度消失/爆炸、数据样本不足导致过拟合、训练速度慢等问题，加之 SVM、LR 等浅层学习算法的优良表现，使得网络的层数无法提升上去，一直局限在几层以内。直到 Hinton 提出一种贪婪无监督逐层学习方法^[25]，才打破了深层神经网络难以训练的僵局，使得神经网络的层数不再受到严重限制，研究人员可以自由的选择神经网络的层数，打开了深度学习的大门。此后，深度学习得到了快速的发展，在视频、文本、图像等多个领域都得到了广泛的研究与应用，成为了机器学习领域的一个热门方法。

随着深度学习的发展，许多著名的算法/结构被相继提出，例如：卷积神经网络 CNN^{[30][31]}、深度玻尔兹曼机^{[32][33]}、多层反馈递归神经网络(RNN)^[34]等。其中卷积神经网络 CNN 网络是一种多层前馈神经网络，如上图 2-2 中 VGG16 就是一种经典的卷积神经网络。CNN 网络在 2012 年 ImageNet 竞赛中的表现极其出色，从此奠定了它在图像识别领域的重要地位，成为了图像领域的核心算法之一^[35]。标准 CNN 网络的模型结构包含输入层、卷积层、池化层、输出层。输入层一般为数

据集中 2 维或 3 维图像；卷积层是图像特征的抽取层，包含一个或者多个卷积核，主要作用为自动提取输入图像对应的图像特征；池化层一般紧接着卷积层出现，对卷积层提取到的特征进行进一步压缩，减小特征图尺寸的同时提取图像的主要特征；输出层出现在最后一个卷积层或者池化层之后，由多个全连接层构成，对卷积层提取到的图像特征进行整合，输出任务所需的向量。

CNN 网络的广泛应用不仅是因为能自动的提取图像对应的图像特征，不再需要研究人员对数据集进行复杂的预处理操作，还因为其具有局部连接和权值共享的特点，这两个特点使得 CNN 网络不仅大幅度降低模型的参数量，缩短了模型的训练时间，而且提高了容错能力，还使模型具有平移不变性。与此同时，计算机 GPU 对 CNN 网络计算的加速效果也为其在图像领域中被广泛的应用提供了巨大帮助，使其成为了图像领域的一种基础结构。经过广大研究人员的开发研究，卷积神经网络已经提出了许多经典的结构，如 AlexNet、VGGNet、ResNet 等。这些结构不仅在 ImageNet 等竞赛中表现出色，而且还常通过迁移学习的方法作为新模型的图像特征提取层，被广泛用于图像领域的各种任务之中，为我们的模型设计与训练提供巨大帮助。

2.2 生成式对抗网络

2.2.1 生成式对抗网络原理

生成式对抗网络（(Generative adversarial network, GAN）是 Goodfellow 等^[36]在 2014 年 6 月提出的一种生成模型，由于其独特的思想和优秀的性能，一经提出就得到了广泛的关注。GAN 模型常与卷积神经网络结合，在图像领域特别是图像生成、图像转换、图像修复等应用中，模型生成结果表现均优于传统生成模型，是当前图像领域中一种新兴的深度学习模型框架。虽然 GAN 模型提出的时间较晚，但是其发展迅猛，到 2018 年 9 月为止，在 GitHub 上 Avinash Hindupur 就统计出了 502 种 GAN 的改进变体结构及应用，GAN 网络的真实数据应该远不止于此。与其它的生成模型不同，GAN 模型由一个生成器和一个判别器共同构成，生成器和判别器之间是一种博弈对抗的关系，通过对生成器和判别器的分别优化训练，最终达到博弈论中的纳什均衡^{[38][39]}。GAN 模型如图 2-3 所示，生成器主要作用是从训

训练数据集中学习数据的真实分布，然后生成器利用输入的随机噪声 z 生成接近数据集分布的生成数据 $G(X)$ ，判别器则需要通过优化自身参数尽可能从输入中分辨出生成器生成的数据 $G(X)$ 和数据集中的真实数据 x ，并输出输入数据来源于数据集的概率。

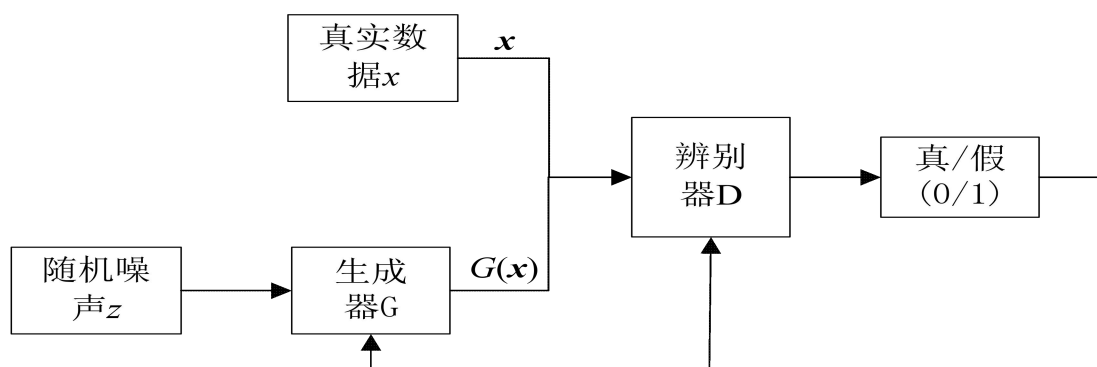


图 2-3 GAN 模型图

GAN 的目标函数为：

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

GAN 网络训练时需要根据目标函数交替优化生成器与判别器，当优化生成器时，最小化目标函数 $V(D, G)$ ，使生成器生成的数据 $G(z)$ 愈加接近数据集真实数据的分布，经过判别器后的输出 $D(G(z))$ 越来越接近于 1，即判别器相同的情况下，优化生成器，使得判别器无法辨别生成数据 $G(z)$ 和真实数据 x 之间的差异，生成数据在判别器眼中达到以假乱真。优化判别器时，最大化目标函数 $V(D, G)$ ，使得生成数据 $G(z)$ 对应的判别器输出 $D(G(z))$ 接近于 0，与此同时真实数据 x 对应的输出 $D(x)$ 尽可能接近于 1，即让生成数据 $G(z)$ 和真实数据 x 通过判别器后的输出值差距尽可能大，使判别器尽可能准确辨别出真实数据 x 和生成数据 $G(z)$ 。生成器要生成尽可能逼真的数据来欺骗判别器，而判别器要通过优化自身参数尽可能的辨别出来自于生成器的生成数据 $G(z)$ ，通过对生成器 G 和判别器 D 依次进行多次优化，分别提升其性能，使的网络生成器 G 与判别器 D 的性能达到纳什均衡，最终使得生成器的生成数据 $G(z)$ 与真实数据 x 足够相似，近似于服从同一个分布，判别器也无法准确识别生成数据和真实数据之间的差异。

2.1.2 生成式对抗网络的改进模型

与其它生成模型相比，GAN 网络拥有许多优点，如：不需要利用马尔科夫链进行反复采样；不需要设计遵循任何种类的因式分解；很多时候实际生成效果优于其它模型等。生成式对抗网络在取得优异生成效果的同时也暴露出了许多问题，主要为：1、模型生成结果不可控。和以往的生成式模型不同，GAN 网络不再对数据集的数据分布进行假设，直接对数据集进行采样，这样的好处是让 GAN 网络的生成结果在理论上可以完全接近数据集的分布，但是模型没有经过预先建模可能会使得模型的生成结果缺乏相应的约束，导致生成结果过于自由，不可控。2、GAN 网络不收敛问题。理论上当 GAN 网络的生成器性能和辨别器性能相当且各自性能都表现很好时，GAN 网络才达到纳什均衡，有较好的生成结果，但是这一个平衡在训练时是十分难以把控的，GAN 网络生成器和辨别器性能经常会失衡，导致模型不收敛。3、生成结果崩溃问题。GAN 网络的没有损失函数，其训练的目标是一个 minmax 问题，因此在进行网络参数学习时我们无法通过目标函数来确定生成器和辨别器的性能在训练过程中是否有提高，这样在 GAN 训练过程中有可能出现生成器与辨别器均不再优化的情况，造成生成结果崩溃的问题，生成器总是生成相同的数据，并且我们难以在训练时及时发现予以调整。

针对模型不可控问题，Mirza 等^[39]基于原始 GAN 提出 GAN 模型的改进结构--条件生成对抗网络(CGAN)，在生成器输入噪声 z 的同时输入一个条件 c ，并且将这个条件 c 也作为辨别器的输入，利用条件 c 对 GAN 的生成结果进行限制。CGAN 模型如图 2-4。

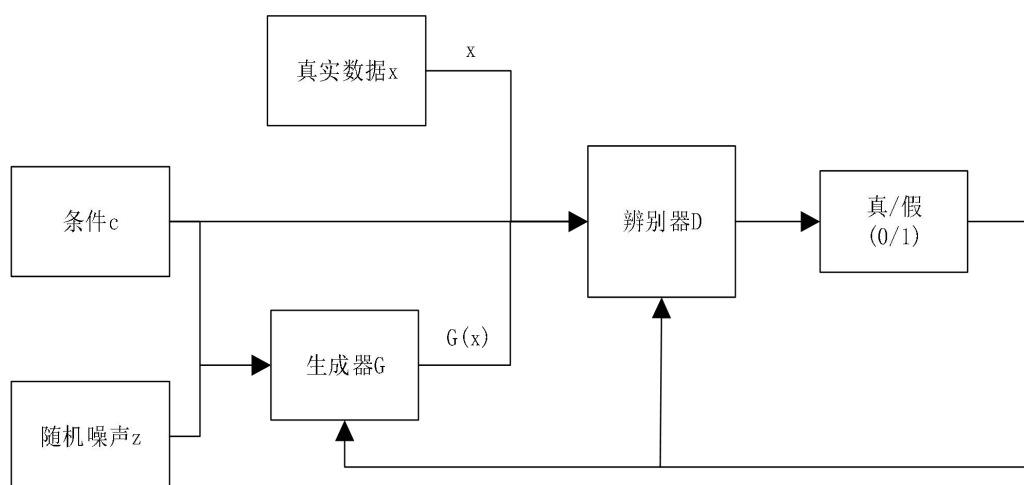


图 2-4 CGAN 网络

CGAN 目标函数 $V(D,G)$ ，如公式(2)所示：

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

针对 GAN 模型不收敛问题和崩溃问题，Arjovsky 等^[40]提出 Wasserstein-GAN (WGAN)，主要对原始 GAN 的训练过程进行改进，具体改进措施如下：1、对于辨别器输出层去除激活函数 2、GAN 的优化目标函数公式中去除 \log 3、辨别器进行参数更新后，将权重强制截取到一个固定范围 4、不使用基于动量的优化算法。通过 WGAN 的这些改进，彻底的解决了原始 GAN 的不收敛问题和崩溃问题，减小了 GAN 网络的训练难度，也使得生成结果的多样性得到保障。WGAN 网络虽然减小了 GAN 网络的训练难度，解决了不收敛问题和崩溃问题，但是 WGAN 网络对权重进行强制截取容易导致模型产生梯度消失或者梯度爆炸问题。对此，Gulrajani 等^[41]提出 Improved WGAN (WGAN-GP)，对 WGAN 模型进一步改进，使用梯度惩罚代替强制截取梯度。Improved WGAN 网络的目标函数为：

$$\min_G \max_D = E_{x \sim P_{data}} [D(x)] - E_{z \sim P_z} [D(G(z))]$$

$$+ \lambda E_{\tilde{x} \sim P_{\tilde{x}}} [(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2] \quad (2-3)$$

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

2.3 词向量模型

由于神经网络无法直接处理文本数据，所以需要对文本数据进行数值转换。传统的转换方法是将被文本数据转换成 one-hot 词向量，one-hot 词向量使用一个维度与词汇表大小相等的多维向量来表示文本数据，在词汇表里每一个词分别与多维向量中的一个维度相对应，当单词存在时向量在对应维度的取值为 1，否则取值 0。例如当前词汇表的大小为 4，词“dog”在对应 4 维向量中的第 0 维，那么它的词向量就是[1, 0, 0, 0]。同样的道理，词“cat”是对应 4 维向量中的第 3 维，词向量就是[0, 0, 0, 1]。one-hot 词向量是一种高维稀疏的词向量转换方法，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，无法体现词与词之间的语义关系。

2013 年 Google 开源一款新词向量生成工具 Word2vec^[42]可以将文本映射成为一个固定的多维空间向量，如 cat 可能被表示为 6 维空间向量[0.2, 0.25, 0.3, 0.01, 0.9, 0.6]，也可能表示为 3 维空间向量[0.5, 0.1, 0.4]，这个多维空间向量的维数与词汇量无关，完全由使用人员在训练时自己决定，目前 Word2vec 被大量应用于自然语言处理(NLP)任务当中。Word2vec 模型的主要思想是具有相同或相似上下文的词汇，可能具有相似的语义，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的 N 维词汇空间，并使语义上相近的词汇在该空间中的距离相近，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，避免 one-hot 词向量的缺点。

Word2vec 模型是一种简单的神经网络模型，其只包含一个隐含层，有 CBOW^[43]和 Skip-Gram^[44]两种训练模式。CBOW 与 Skip-Gram 模型如图 2-5 所示，CBOW 模型输入层为需要预测的单词 $w(t)$ 的上下文 $w(t-2)$ 、 $w(t-1)$ 、 $w(t+1)$ 、 $w(t+2)$ 等，输出层输出为预测的单词 $w(t)$ 。

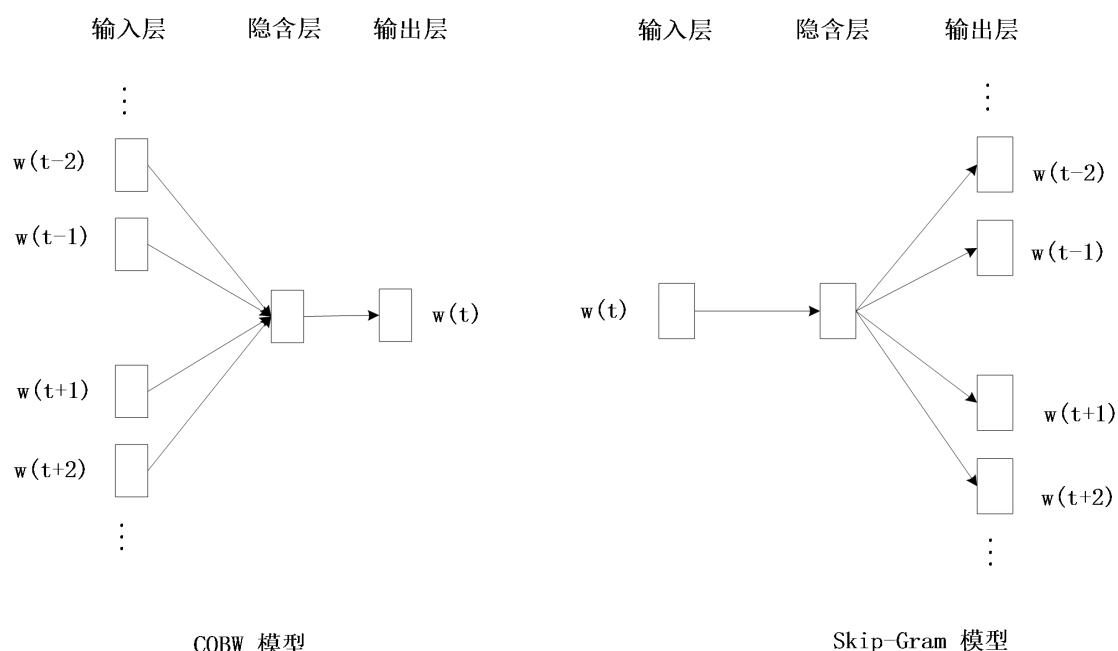


图 2-5 CBOW 与 Skip-Gram 模型

与 CBOW 模型相反，Skip-Gram 模型输入层为单词 $w(t)$ ，输出层的输出为单

词 $w(t)$ 的上下文 $w(t-2)$ 、 $w(t-1)$ 、 $w(t+1)$ 、 $w(t+2)$ 等。通过数据集对模型进行训练后，网络中的隐含层即为所需要的词向量转换矩阵，通过这个转换矩阵可以将词汇表中的词汇转换为一个多维词向量，通过选择隐含层的神经元数目就可以确定 Word2vec 模型的输出向量维数。因此，把训练好的 CBOW/ Skip-Gram 模型的输出层去除，将原隐含层作为模型的输出层得到的新模型就是 Word2vec 模型。当训练数据集的词汇量过大时，通过选择一个神经元数目小于词汇量的隐含层就可以实现对词向量的降维。

2.4 迁移学习

在传统机器学习的方法中，模型要取得良好的性能一般要满足 2 个条件：1、要拥有大量与任务相关的数据。模型需要先从大量数据中提取与任务相关知识，然后才能将学到知识应用到测试任务中，没有大量的数据支撑，模型无法学习到正确的知识，实际测试结果将会表现很差。2、用于训练的数据分布要与测试数据属于独立同分布。只有训练数据和测试数据属于同一分布，才能保证模型在训练集中学习到的知识在测试时同样有效。所以传统机器学习得到的模型通常只适用于当前任务，只要新任务数据与当前有所不同，模型的效果就会变得很差，这意味着只要任务数据发生变化就要重新开始训练模型。在实际应用过程中，虽然当前已经有许多的开源数据集（如 ImageNet、NUS-WIDE、MSRA-MM 等），但是对于大多数的特定应用，开源数据集并不符合应用需求，同时已有与任务相关的数据集通常比较小，而重新收集数据又具有很大的难度，这样使得机器学习在许多特定领域的应用困难重重。

针对机器学习的以上问题，一种新的机器学习方法--迁移学习被提了出来，其示意如下图 2-6。迁移学习(Transfer Learning)目标是将从一个环境中学到的知识应用到新环境中，帮助解决新的学习任务^{[3][45]}，如图像领域中同一种物体的图像具有相同的底层特征（边缘、视觉形状、几何形状等），因此对于一些数据集较小的特定应用，可以将从某一大数据集中学习到的该物体特征用于当前任务中，从而解决数据集过小的问题。

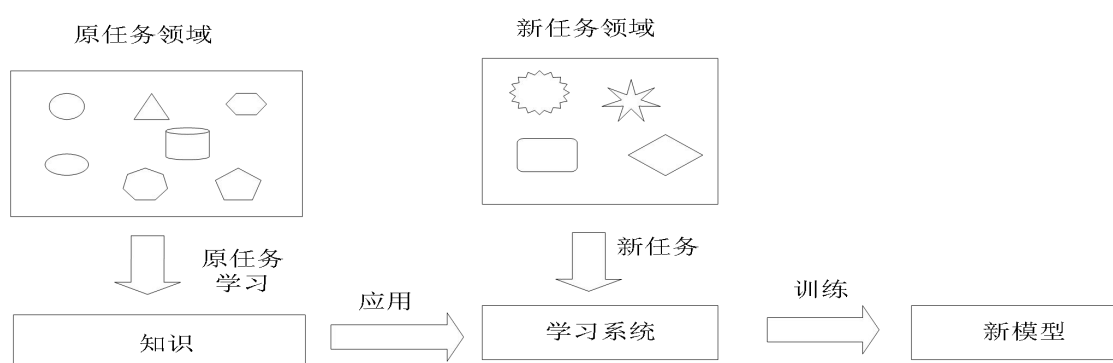


图 2-6 迁移学习

当前常用的迁移学习方式主要为：1、当数据集的分布差异较大时，可以将一些表现比较好的经典神经网络结构（如 VGG16、ResNet 等）或者相似任务中的神经网络结构借用到自己的神经网络模型中，通过修改原模型的输出部分的网络结构使其符合新任务需求，再通过新任务的数据集对网络模型的权重进行重新训练；2、在借用网络结构的同时保存原网络模型的部分或者全部权重，在通过新数据集对整个模型进行训练时，只需对原模型训练的权重进行微调就能很好的解决新任务，这种方法多用于新任务数据集较小的情形。

目前，在图像处理领域中常将一个经过大数据集（一般为 ImageNet 数据集）预训练的经典网络结构（VGG16、ResNet、inception-resnet 等）通过迁移学习的思想，用于新图像处理任务中，作为新任务模型结构的图像特征提取器。这样图像特征提取器的参数固定，所以新任务模型可以不再学习这部分的参数，大大减少了模型的参数量。通过迁移学习的方法可以在实现图像特征提取的同时减小模型对于数据量的需求，有效的解决某些领域数据集数据量不足的问题。

2.5 本章小结

本章主要介绍了本文模型使用的几个关键技术，首先对深度学习和卷积神经网络模型进行简单的介绍，它们是本文的其它技术的基础；其次，分析了本文模型关键结构生成式对抗网络的原理，通过对原始 GAN 缺点的阐述引入本文使用的 CGAN 和 WGAN-GP 两种改进结构；通过对当前 one-hot 词向量缺点的分析，引出本文采用 Word2vec 模型；最后，对本文标注模型中 CNN 特征提取模块使用的迁移学习方法进行了简单介绍。

3 基于 GAN 的标注模型

3.1 标注模型设计思想

基于深度学习和部分基于传统机器学习的标注模型之所以会产生模型输出层神经元数目（或分类器）与数据集中的标注词汇量成比例的问题，是因为以下 2 点原因：1、采用 one-hot 向量对标注词进行向量化，one-hot 向量化方式使得标注对应的词向量维数与数据集词汇量相等 2、模型的输出采用一次输出所有图像对应标注的方式，这种输出方式要求模型的输出结果要覆盖到数据集中的每一个标注，即要覆盖到词向量的每一维度，这使得输出层的神经元或者分类器数目需要等于词向量维数（某些分类器可能是大于词向量维数）。综合以上两点，模型的输出层神经元数目需要与词向量维数相等，也就是与数据集标注词汇量相等。

为了解决模型的输出层神经元数目与图像标注数据集中的标注词汇量成比例的问题，本文从上文的 2 个点原因进行改进：1、使用 Word2vec 模型代替 one-hot 进行标注的向量化。通过 Word2vec 模型进行向量化，标注词向量的维数可以自由选择，不再与标注词汇量相关。2、使用 GAN 网络每次生成一个向量，生成向量的维数与 Word2vec 模型输出词向量维数相同，即每次生成图像的某一个标注对应的词向量，然后 GAN 网络通过随机噪声的扰动作用输出不同向量，实现生成图像对应的不同的标注。

通过以上 2 点改进，使得标注模型的输出层神经元数目只与可自由选择的多维向量维数相关，不再与标注词汇量相关，解决了模型的输出层神经元数目（或分类器）与图像标注数据集中的标注词汇量成比例的问题。与此同时，当标注词汇量发生增删时，只需要对 Word2vec 模型的词向量转换表进行增删，不再需要对生成器 G 和判别器 D 的结构进行修改；当标注词汇量较大时，Word2vec 模型转换的词向量可以一个较低的维数，实现对词向量的降维，使模型的输出层神经元数目不会因词汇量较大而变的太大，利于模型的设计与训练。

3.2 标注模型整体结构

根据前文标注模型的设计思想，对本文的标注模型进行了整体设计实现，模型

的整体结构如下图 3-1，标注模型主要分为数据预处理模块、GAN 网络训练模块与性能测试模块 3 部分。

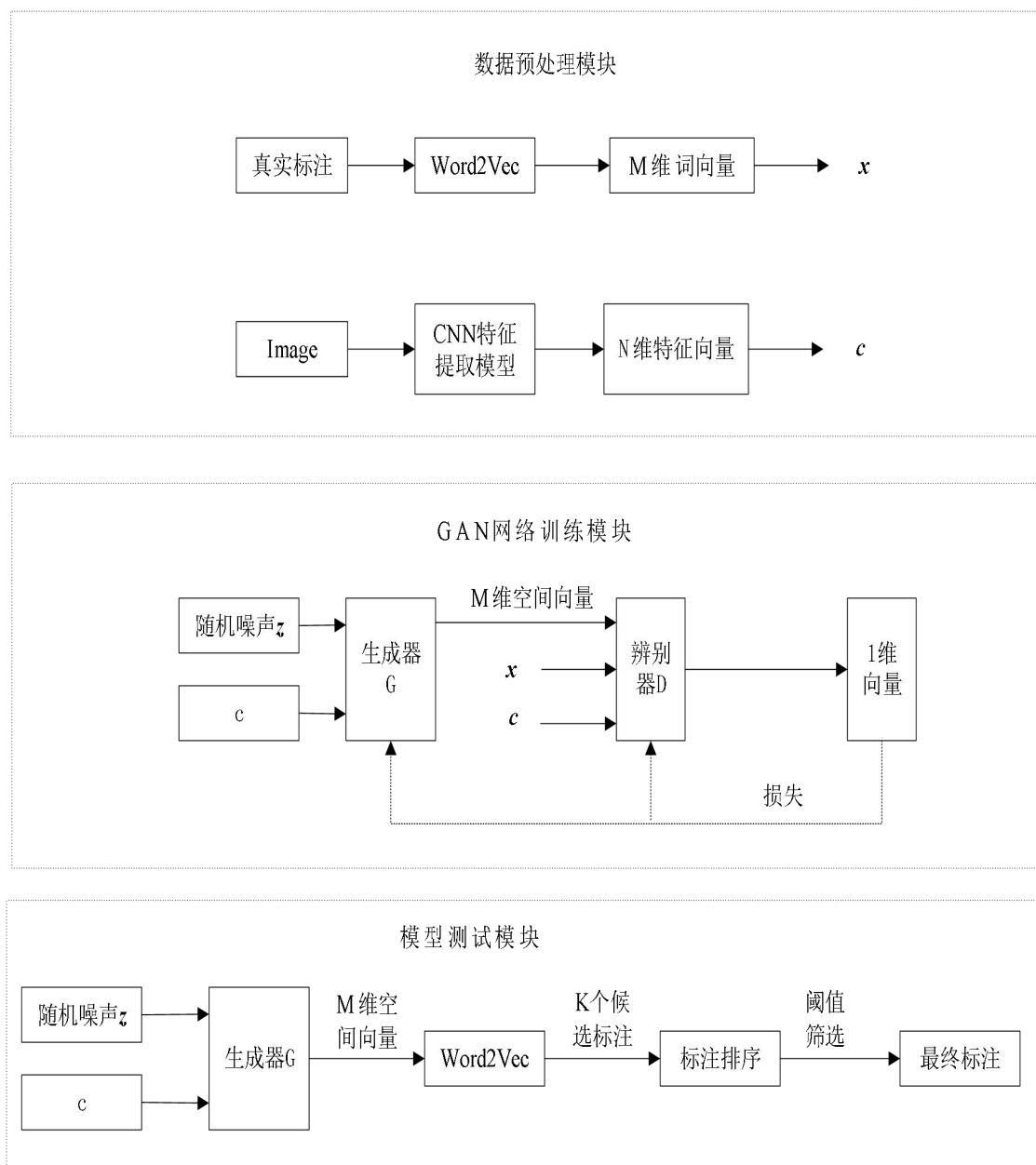


图 3-1 模型整体结构图

数据预处理模块主要功能为：1、对原始数据集图像的尺寸进行调整，使图像大小满足模型的输入要求，对较小的数据集进行数据增广，扩充训练数据的数据量，提高模型的泛化能力 2、对数据集进行改造，将原始数据集中图像与所有标签组成的数据对改成每幅图像与其对应的单个标签组成的数据对，构成符合本文

模型的新数据集 3、利用数据集中的标注对 Word2vec 模型进行训练，利用训练后的 Word2vec 模型对新数据集中的标注进行词向量转换 4、使用迁移学习的方法提取数据集中图像对应的图像特征。

GAN 网络训练模块主要功能是构建 CGAN 学习模型，模型根据训练数据集学习图像特征与标注之间的潜在联系，为测试模块提供学习好的生成器作为测试时的实际标注模型，对未知图像进行标注；模型判别器主要用于辅助对生成器的训练，提高模型生成器的性能。

模型测试模块的主要功能为：1、通过训练好的模型生成器对测试图像进行标注获取生成的词向量，利用 Word2vec 模型获取生成向量对应的标注词，输出图像最终标注 2、根据测试数据集和模型生成器计算出模型的标注性能指标值，为模型的实际标注性能比较提供参考。

3.3 数据预处理

3.3.1 图像数据处理

1、图像大小处理与特征提取模块

在图像数据集中的图像大小一般不固定，如 Corel 5k 数据集中有 128*192、192*128 两种大小的图像，IAPRTC-12 数据集有 320*480、360*480、480*360 等多种尺寸的图像，而神经网络一般要求有固定大小的输入，为了满足模型的输入要求，需要将所有的图像调整到统一的大小。

对于 CNN 图像提取模型，本文采用迁移学习的方法，将在 ImageNet 数据集上进行预训练的 Inception- ResNetV2 模型^[46]去除输出全连接层，保留所有权重数据，作为本文图像提取模型。Inception- ResNetV2 模型是谷歌提出的一种经典卷积神经网络结构，在 Inception 网络的基础上引入了残差结构，提高网络性能的同时还明显加速了网络的训练速度，其在 ImageNet 图像数据集上取得了空前的成功，将 ImageNet 比赛的 Top1 和 Top5 准确率分别提高到了 80.3%和 95.3%。Inception- ResNetV2 模型在 ImageNet 图像数据集的优异表现，说明其模型结构对 ImageNet 数据集图像特征的提取与识别具有较高的准确性，而本文标注数据集和 ImageNet 数据集的图像内容较相似，所以通过迁移学习将其作为本文 CNN 图像提取模型，

一方面可以保证模型的图像特征提取的准确度，另一方面也可以减小模型的参数量，避免因标注数据集过小带来模型过拟合。因为 Inception-ResNetV2 模型的标准输入图像大小为 $299 \times 299 \times 3$ ，所以将数据集中所有图像调整到 $299 \times 299 \times 3$ ，通过 Inception-ResNetV2 模型输出 1536 维图像特征。

2、数据集数据增广

为了使模型具有更强的泛化能力，减小模型的过拟合，这就需要我们为模型提供一个较大的数据集，使数据尽可能多样化，才能避免让模型学习到与任务不相关的特征，降低模型的性能。但在现实情况中，部分的常用的数据集较小，如本文采用的 Corel 5K 数据集只有 5000 张图像数据，而重新收集制作数据集的代价又比较大，所以对原始数据集进行数据增广是十分必要的。常用的图像数据增广的方式主要为：几何变换和像素变换两种。几何变换包括对图像进行水平/垂直翻转、平移、缩放、裁剪等方法，像素变换主要为对图像增加噪声和滤波、改变图像通道顺序、调整对比度等方法。本文采取的数据增广方式主要为：1、对原数据图像进行水平翻转 2、对图像的像素值进行小幅度的增减 10 个像素值 3、给图像随机增添高斯噪声 4、给图像随机增加椒盐噪声。

本文数据增广的效果图如下图 3-2，从图中可以看出，数据增广的每种方法都可以得一幅到与原图对应的新图像。这些新图像中的物体特征与原图像相同，对于人眼而言图像的改变很小，甚至可以忽略，并不影响图像的内容识别，但是由于神经网络识别的是图像的像素值，所以对于模型而言，这些增广得到的新图像在像素值上与原图像具有很大差异，可以视作为不同图像。通过数据增广的方式将 Corel 5K 数据集的大小提高很多倍可以满足模型学习的需求，提高模型的泛化能力。



图 3-2 数据增广

3.3.2 标签数据处理

为了解决输出层神经元数目与标注词汇量相关的问题，本文模型采取 Word2vec 模型对标注进行向量化。由于 Word2vec 模型训练的数据一般要求具有上下文关系的语句或者是文章，而数据集中的图像标签之间是相互独立的，并没有如一般语句中的上下文关系，所以将每幅图像对应的所有标注组合作为一组训练语句。这样通过数据集对 Word2vec 模型训练之后，不仅可以获取标注对应的多维词向量，还可以使共现次数较高的标签之间的词向量距离较近。本文的 Word2vec 模型采用 gensim 函数库的 Word2vec 模块实现，Word2vec 模块不仅提供 Word2vec 功能实现的接口，还实现了反向查找向量对应标注、标注间距离计算、增量训练等功能。

在原始数据集中，图像与其对应的所有标注组成一组数据，但是由于本文模型每次只输出一个图像对应的标注，所以需要处理原数据集，将数据集中图像与每一个标签分别组成一组新数据，重构模型的数据集，然后再利用 Word2vec

模型将数据集中的标签转换成对应的多维词向量，作为模型的真实数据 x 输入到模型辨别器中。

3.4 GAN 网络训练模型

训练模型整体采用 CGAN 架构，由生成器和辨别器构成，CGAN 模型的限制条件 c 为 1536 维图像特征，这个图像特征是由数据预处理模块中 CNN 图像特征提取器输出，即是由迁移学习处理的 Inception-ResNetV2 模型的输出。在模型具体网络的设计上，生成器和辨别器的实现与常用卷积神经网络输出层相同，整体上均采用多个全连接层相连的方式构建。全连接层可以实现对输入向量所包含的信息提取与整合，通过多个全连接层相连，生成器可以深入提取图像特征和随机噪声中包含的信息，生成与图像相关的标注词向量；辨别器将图像特征与输入词向量信息进行深入整合，辨别输入词向量来源于数据集还是生成器，输出一个 1 维向量，用于构建模型的损失。

在生成器和辨别器中除了采用全连接层以外，还使用批标准化层、激活层（激活函数）、Dropout 层来辅助模型的训练。模型中全连接层用于对输入向量的信息进行进一步整合和挖掘，批标准化层紧接着全连接层，将全连接层的输出向量强行调整为均值为 0 方差为 1 的标准正态分布，使得在网络训练过程中每个全连接层的输入保持相同分布的，这样不仅可以加快模型的收敛速度，减少训练时间，还可以降低网络对初始化权重的敏感性。激活函数紧接着批标准化层，为模型引入非线性，提高模型的拟合能力，本文激活函数采用 leaky_relu 函数，leaky_relu 函数在保留 relu 激活函数收敛速度快、计算速度快优点的同时避免 relu 激活函数会导致神经元死亡的缺点。Dropout 层是对经过激活函数后的输出向量的每一维度进行随机归 0 操作，随机归 0 的概率一般采用 50%，通过 Dropout 层可以减小模型的过拟合，在模型的最后几层不采用 Dropout 层以防止信息丢失。

3.4.1 生成器模型结构

生成器输入为 100 维的随机噪声 z 和 1536 维图像特征（作为条件向量 c ），输出 500 维生成词向量（与 Word2vec 模型输出词向量维数相同），其模型如下图 3-3 所示。首先，100 维随机噪声通过全连接层 1 输出 512 维向量，1536 维图像特征

经过全连接层 2 输出 2048 维向量，两个向量拼接得到 2548 维向量，全连接层 3 和全连接层 4 重复这一过程得到 2548 维向量；其次，2548 维向量通过全连接层 5 输出 1024 维向量，再通过全连接层 6 输出本文所需的 500 维生成词向量。生成器的输出层采用 \tanh 作为激活函数，使得输出的生成词向量取值范围保持在 $[-1,1]$ ，以保证生成词向量不仅在维数上还是在取值范围上均与 Word2vec 模型输出标注对应的词向量相同。

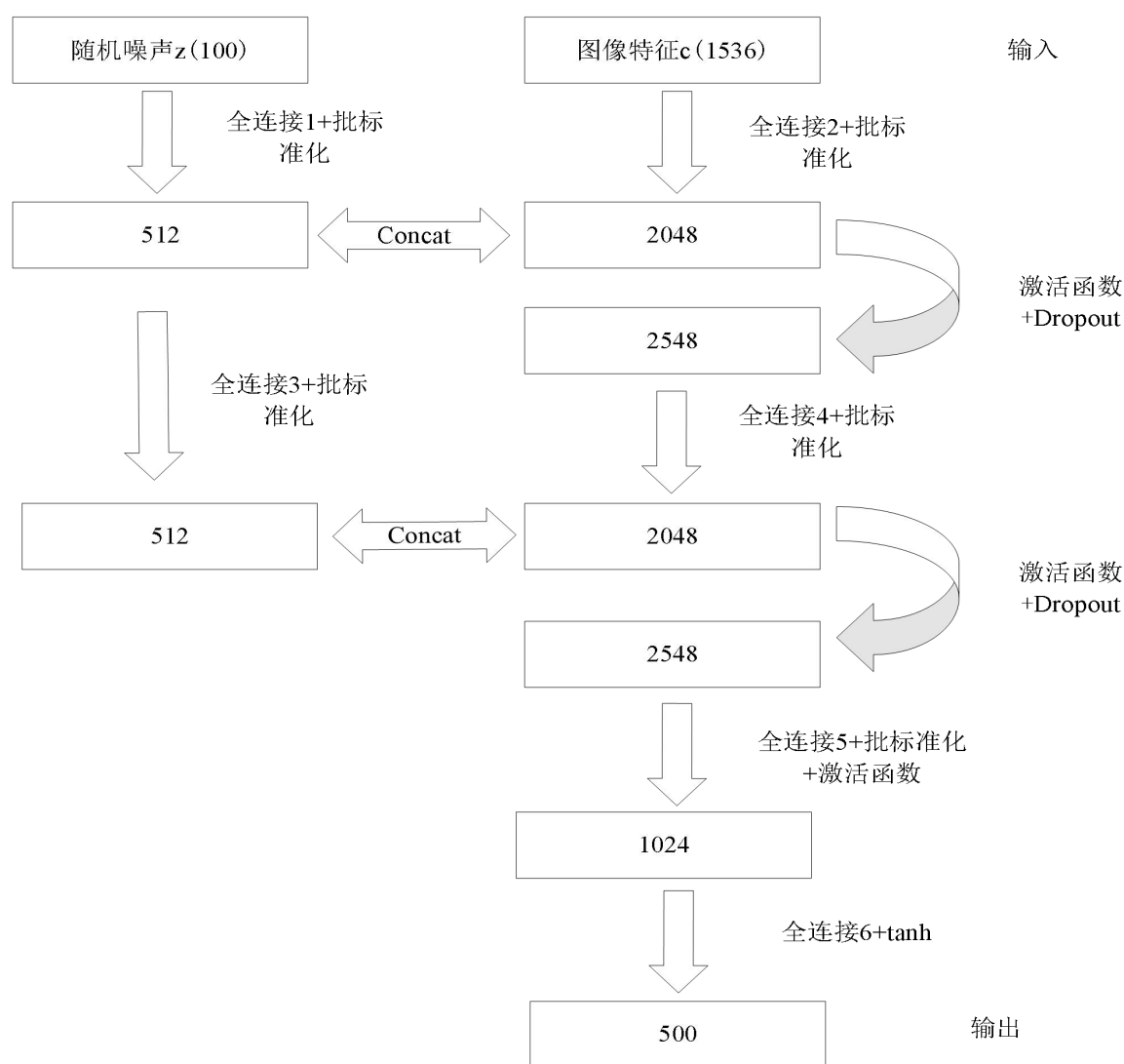


图 3-3 生成器结构

3.4.2 判别器模型结构

判别器的模型结构如图 3-4，判别器除了输入 1536 维条件向量，还输入经 Word2vec 模型转换的 500 维标签词向量或者 500 维生成器生成的词向量，输出对

应的 1 维向量，作为模型损失构建的基础。首先，500 维词向量通过全连接层 1 输出 1024 维向量，1536 维图像特征经过全连接层 2 输出 2048 维向量，两个向量拼接得到 3072 维向量，全连接层 3 和全连接层 4 重复这一过程得到 3072 维向量，最终通过全连接层 5 和全连接层 6 输出 1536 维向量；其次，1536 维向量通过通过全连接层 7 输出 512 维向量，最后通过全连接层 8 输出 1 维向量用于构建模型损失。由于本文模型将 WGAN-GP 模型引入到本文模型中，所以判别器的模型结构中不使用批标准化，输出层也不使用激活函数，其余层激活函数采用 leaky_relu 激活函数。

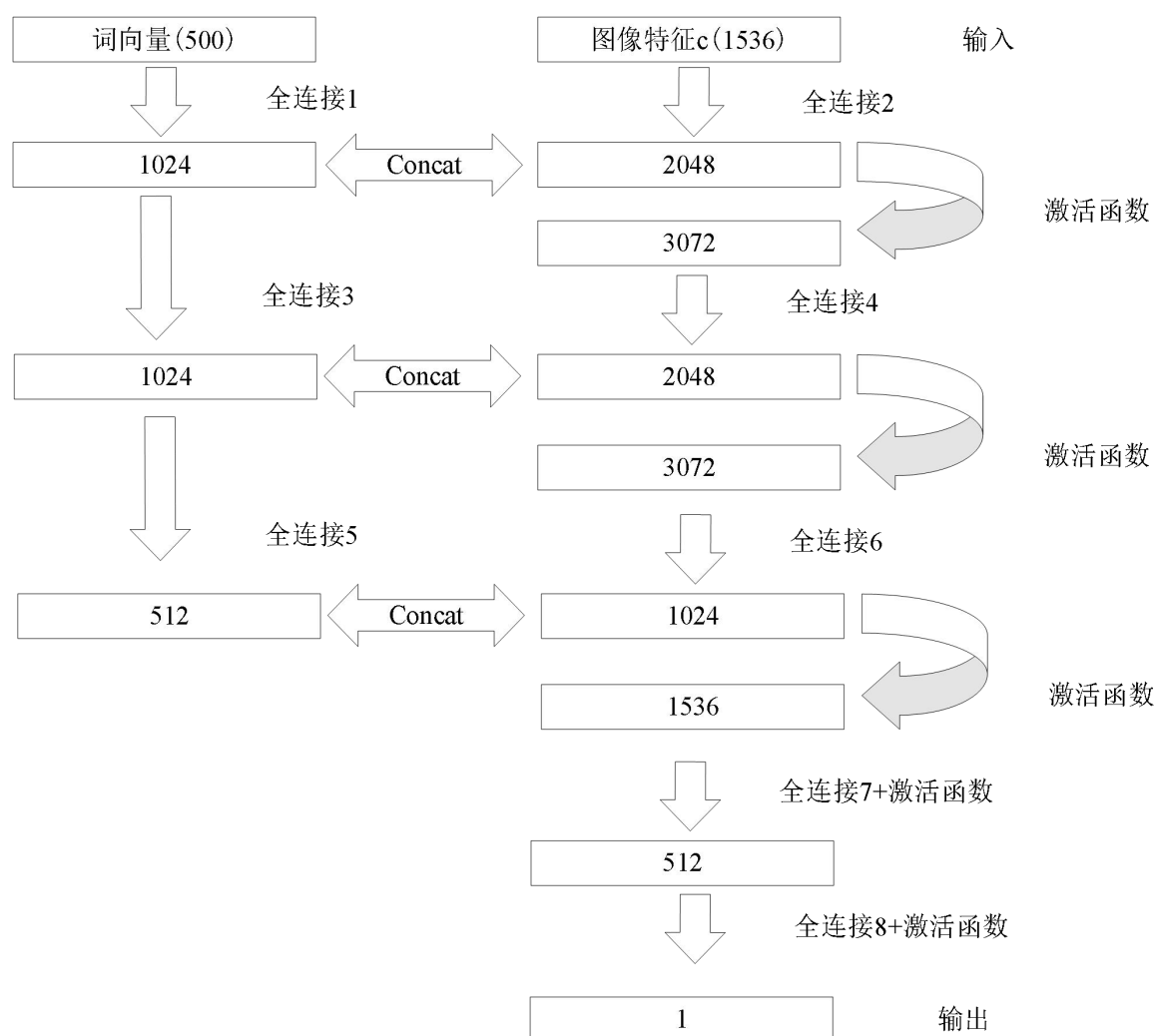


图 3-4 判别器结构

3.5 模型损失

本文模型的损失函数由生成对抗损失、L2 损失及感知损失构成。为了解决原始 GAN 模型训练难、易崩溃的问题,模型将 WGAN-GP 模型引入到模型训练之中,所以本文使用 WGAN 模型的损失函数作为本文生成对抗损失,来衡量生成词向量与真实标签对应词向量之间的整体差异。假设判别器输出为 $D(x)$, 生成词向量为 X_G , 真实标签对应词向量为 X , 则生成器和判别器对应的生成对抗损失 $L_{GAN}(G)$ 和 $L_{GAN}(D)$ 分别为:

$$L_{GAN}(G) = -D(X_G) \quad (3-1)$$

$$L_{GAN}(D) = D(X_G) - D(X) + \lambda(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2 \quad (3-2)$$

其中, \tilde{x} 为真实标注词向量 X 与生成器生成词向量的插值, 计算如下:

$$\tilde{x} = X + k*(X_G - X), k \in [0,1] \quad (3-3)$$

如同 GAN 网络在图像生成领域, 生成对抗损失使得生成图像与原数据集图像整体相似, 但是许多细节不同一样, 本文模型中生成对抗损失使得生成词向量与真实标签对应词向量之间整体相似, 但在某些维度上取值可能差异较大, 这样会使得模型的某些标注变得不准确, 影响模型的性能。因此, 在损失函数中引入 L2 损失来衡量生成词向量与真实标签词向量在每一维度上的差异, 使得生成向量在整体上和每一维度上都比较接近。L2 损失即均方误差, 表示生成词向量与标签词向量之间距离的平方和, 模型 L2 损失 $L_{L2}(G)$ 计算如下:

$$L_{L2}(G) = \frac{1}{n} \sum_{i=0}^{n-1} (X^i - X_G^i)^2 \quad (3-4)$$

其中, 上标 i 表示向量的第 i 维度。

除了通过生成对抗损失和 L2 损失使生成词向量与真实标签词向量之间在整体上和每一维度上相似之外, 本文还引入感知损失对网络的标注性能进行进一步提升。与大多数损失不同, 感知损失并不是用模型输出值来构建, 而是采用模型中隐藏层的输出来构建。本文感知损失 L_P 选取的隐藏层为判别器中靠近输出层的 3 个隐藏层, 利用隐藏层输出差异的绝对值来构成感知损失, 以此来进一步提高模型在高级抽象层上的相似度, 进一步减小生成词向量与真实标签词向量之间的差

异。假设 $D^i(x)$ 代表判别器中第 i 层的输出，感知损失 L_p 计算如下：

$$L_p = \sum_i |D^i(X) - D^i(X_G)| \quad (3-5)$$

另外，标注词汇在数据集中的分布不均匀是一个常见的问题，有些标注如 `cafe`、`butterfly` 在 Corel5K 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。由于数据集中不同标注之间的词频差异巨大，如果不进行处理，模型容易忽略低频标注的影响，导致对低频词汇标注的准确率和召回率下降，影响模型性能。针对标注分布不均衡问题，本文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，使得词频低的标注在一个训练批次的损失中具有更大权重，词频较高的标注损失占比较小。最终，模型的整体损失由生成对抗损失 L_{GAN} 、L2 损失 L_{L2} 及感知损失 L_p 共同加权构成，生成器损失 L_G 和判别器损失 L_D 计算如下：

$$L_G = \alpha[\lambda_1 L_{GAN}(G) + \lambda_2 L_{L2}(G) + \lambda_3 L_p] \quad (3-6)$$

$$L_D = \alpha[\lambda_4 L_{GAN}(D) + \lambda_5 L_p] \quad (3-7)$$

其中 $\alpha = 10/m$ ，为词频的平衡系数， m 为标签对应的图像数目， λ_1 到 λ_5 表示各个损失对应的权重。

3.6 测试模块

由于本文模型的生成器每次只输出一个候选标注对应的词向量，所以为了获取准确而且全面的标注结果，需要模型对测试图像进行足够多次数的预测，然后对所有的预测结果进行排序确定最终标注。具体测试过程为：1. 利用已训练好的生成器模型，输入 100 维随机噪声 z 和测试图像对应的图像特征，进行 N 次预测，其中 N 要足够大，本文预测次数选择一个 `batch_size`（128 次），每次预测得到一个候选标注词向量，总共获得 N 个候选词向量 2. 对于每个候选词向量 X_G ，使用 Word2vec 模型的 `similar_by_vector` 函数，反向查询数据集标注词汇中与其距离最近的 M 个标注，构成候选标注集合 W ，其中 M 值可以自由选择，并记录候选词向量 X_G 与该标注匹配的概率 p 3. 对每个标注的概率 p 进行叠加作为该标注的出现次数 m ，对所有标注的出现次数进行统计排序，最终通过一个统一的阈值超参

数 k 筛选出出现次数大于阈值 k 的标签作为该测试图像的最终标注。

对于每幅测试图像，通过将模型的最终标注结果与原始数据集的真实标注进行比较可以获得模型对单幅图像的标注性能，对所有测试图像的标注性能进行平均即可得到模型的最终标注性能。模型标注/测试的基本流程，如下：

Algorithm 1: 模型标注测试算法

输入: 条件向量 c (图像特征向量), 随机噪声 z , 预测次数 N

输出: 图像对应标注

```
1:  for  $i = 0$  to  $N$  do
2:      生成器根据条件  $c$  和随机噪声  $z$  生成词向量  $X_G$ 
3:      Word2vec 模型根据  $X_G$  获取  $M$  个最接近标签构成集合  $W$  并记录每个标签与词向量匹配的
        概率  $P$ 
4:  end for
5:  for each label  $\in W$  do
6:      对  $N$  次预测结果中的每个标签 label 对应的概率分别进行叠加, 得到每个 label 对应的
        出现次数
7:  end for
8:  根据标签 label 的出现次数对 label 进行排序
9:  选择出现次数大于阈值  $K$  的标签作为图像对应标注 (最多为 5 个)
10: 将模型的标注结果与原始数据集的标注结果进行对比, 得到模型对当前图像的标注性能
```

3.7 GAN 模型训练算法

GAN 模型的训练方法为依次对判别器和生成器进行训练, 计算模型损失, 优化自身参数。每轮训练时判别器更新 k 次 (一般为 5), 生成器更新一轮。在训练判别器时, 首先, 利用训练好的 Word2vec 模型将图像对应的标注转换为一个固定维数的多维空间向量, 作为模型的真实数据 X , 多维空间向量的维数自由选择, 可以根据标注词汇量的大小、数据集大小、模型设计难度等进行调整。数据集中的图像经过 CNN 特征提取模块输出对应的图像特征, 作为标注模型的条件 c 。其次, 生成器根据随机噪声 z 和图像特征向量生成一个图像对应的标注词向量 X_G 。由于模型存在随机噪声 z 的扰动, 所以对于相同的条件 c (图像特征), 生成器也可能输出不同的词向量, 即对同一幅图像生成不同的标注词向量。对于每幅图像,

通过生成向量 X_G 和真实数据 X 的插值操作得到插值向量 \tilde{x} ，用于后续模型梯度惩罚损失计算。最后，判别器 D 以输入的条件 c 作为依据，对依次输入的生成向量 X_G 、真实标注词向量 X 、插值向量 \tilde{x} 进行判别，分别输出一个 1 维向量。根据输出的 1 维向量计算判别器损失，从而对判别器自身参数进行优化，完成 1 次判别器的训练过程。重复这一过程 K 次就完成一轮模型训练中的判别器训练过程。对于模型生成器的训练，在判别器训练完成后，生成器根据随机噪声 z 和条件 c 再次生成词向量 X_G ，判别器随之输出 $D(X_G)$ ，生成器根据 X 、 X_G 、 $D(X_G)$ 构建生成器损失，优化生成器参数，完成生成器的训练，也就完成模型的一轮参数学习过程。

通过模型的多轮训练，交替构建判别器和生成器损失分别对判别器和生成器的参数进行优化，提高生成器和判别器性能，使模型的标注更加准确，最终达到纳什平衡。模型的训练算法如下：

Algorithm 2: GAN 模型训练算法

输入：训练总批次 n ，随机噪声 z ，每轮训练判别器更新次数 k ，批尺寸大小 $batch_size$

输出：

```
1: while epoch < n do
2:   for i = 0 to k do
3:     从数据集中抽取 1 个  $batch\_size$  的训练图像，分别提取对应的图像特征作为模型
       训练的条件  $c$ 
4:     每幅图像抽取 1 个标注通过 Word2vec 模型转换成词向量，作为模型真实数据  $X$ 
5:     生成器根据随机噪声  $z$  和条件  $c$  生成词向量  $X_G$ 
6:     通过对真实数据  $X$  和词向量  $X_G$  插值得到向量  $\tilde{x}$ 
7:     判别器根据条件  $c$  分别对真实数据  $X$ 、生成词向量  $X_G$ 、插值向量  $\tilde{x}$  分别输出
        $D(X)$ 、 $D(X_G)$ 、 $D(\tilde{x})$ 
8:     根据  $D(X)$ 、 $D(X_G)$ 、 $D(\tilde{x})$  计算判别器损失  $L_D$ ，优化判别器参数
9:   end for
10:  生成器根据随机噪声  $z$  和条件  $c$  再次生成词向量  $X_G$ ，判别器输出  $D(X_G)$ 
11:  根据  $X$ 、 $X_G$ 、 $D(X_G)$  等计算生成器损失  $L_G$ ，优化生成器参数
12: end while
```

3.8 本章小结

本章主要针对前文分析的输出层神经元数目与标注词汇量相关的问题设计本文的新标注模型，模型通过 Word2vec 模型将标注转换为一个固定的多维词向量，再利用生成式对抗网络生成图像对应的同一维度的标注词向量，达到将输出层神经元数目与标注词汇量解绑的目的。本章从模型的数据预处理方式、训练网络结构、损失函数设置、测试方法等多个方面对模型进行详细的介绍，简述模型的工作原理并给出模型的训练、测试的算法流程。

4 模型实验及结果分析

4.1 实验数据集

本文实验的数据集为图像标注领域常用数据集：Corel 5K 数据集^[47]和 IAPRTC-12 数据集^[48]。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之一。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像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

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

4.2 评估方法

本文实验采用的评价方法是图像标注工作中常用的评价指标：准确率(P)和召回率(R)及 F1 值。对于标签 x ，假设在测试集中被标签 x 标注的图像数目为 N ，模型对测试集中所有图像进行标注后，在所有标注结果中被标签 x 标注的图像总数为 N_1 ，在被标签 x 标注的 N_1 张图像中标注正确的图像数量为 N_2 。那么，对于标签 x ，模型标注的准确率(P)和召回率(R)及 F1 值计算如下：

$$P = N_2 / N_1$$

$$R = N_2 / N \quad (4-1)$$

$$F1 = 2 * P * R / (P + R)$$

整个模型标注的准确率(P)和召回率(R)及 F1 值为所有测试集标签的准确率(P)和召

回率(R)及 F1 值的平均值。

4.3 Word2vec 模型训练参数对模型标注性能的影响

Word2vec 模型是本文标注模型的一个重要组成部分，而 Word2vec 模型是一种无监督的学习方式，所以没有明确的指标来直观的评判 Word2vec 模型训练结果的好坏。为了探究不同训练参数下的 Word2vec 模型对模型标注性能的影响，给后续实验选择一组相对较好的 Word2vec 模型训练参数，本文选择 Word2vec 模型 3 组重要的训练参数在 Corel 5K 数据集上进行实验，所选参数分别为：模型训练次数 iter、高频词汇采样频率 sample、Word2vec 训练模型 sg 及优化方式 hs。其中 sg 取 1 表示采用 Skip-Gram 模型，取 0 表示采用 CBOW 模型训练，hs 为 1 表示模型基于 hierarchical softmax 优化，hs 为 0 表示模型基于 Negative Sampling 优化。为了减少调参过程对模型标注性能带来的影响，模型每次训练时只改变 Word2vec 模型的 3 组参数，并且模型开始训练后所有参数不再进行调整。由于每个参数的取值范围较广，为了使实验结果曲线彼此之间能清晰分辨，所以仅选取 6 组有代表性 Word2vec 模型训练参数进行标注性能曲线绘制，下图 4.1 实验结果曲线对应的 Word2vec 模型参数设置如下表：

表 2 Word2vec 模型参数设置表

曲线	iter	sample	hs	sg
1	500	0.001	0	0
2	100	0.001	1	1
3	500	0.001	1	1
4	500	0.001	0	1
5	500	0.001	1	0
6	500	0.01	1	1

图 4.1 为不同 Word2vec 训练参数下的模型标注性能图，模型选择能同时反映准确率 P 和召回率 R 的 F1 值作为对比的参考指标，模型训练批次(epoch)为 150，此时模型性能基本稳定。

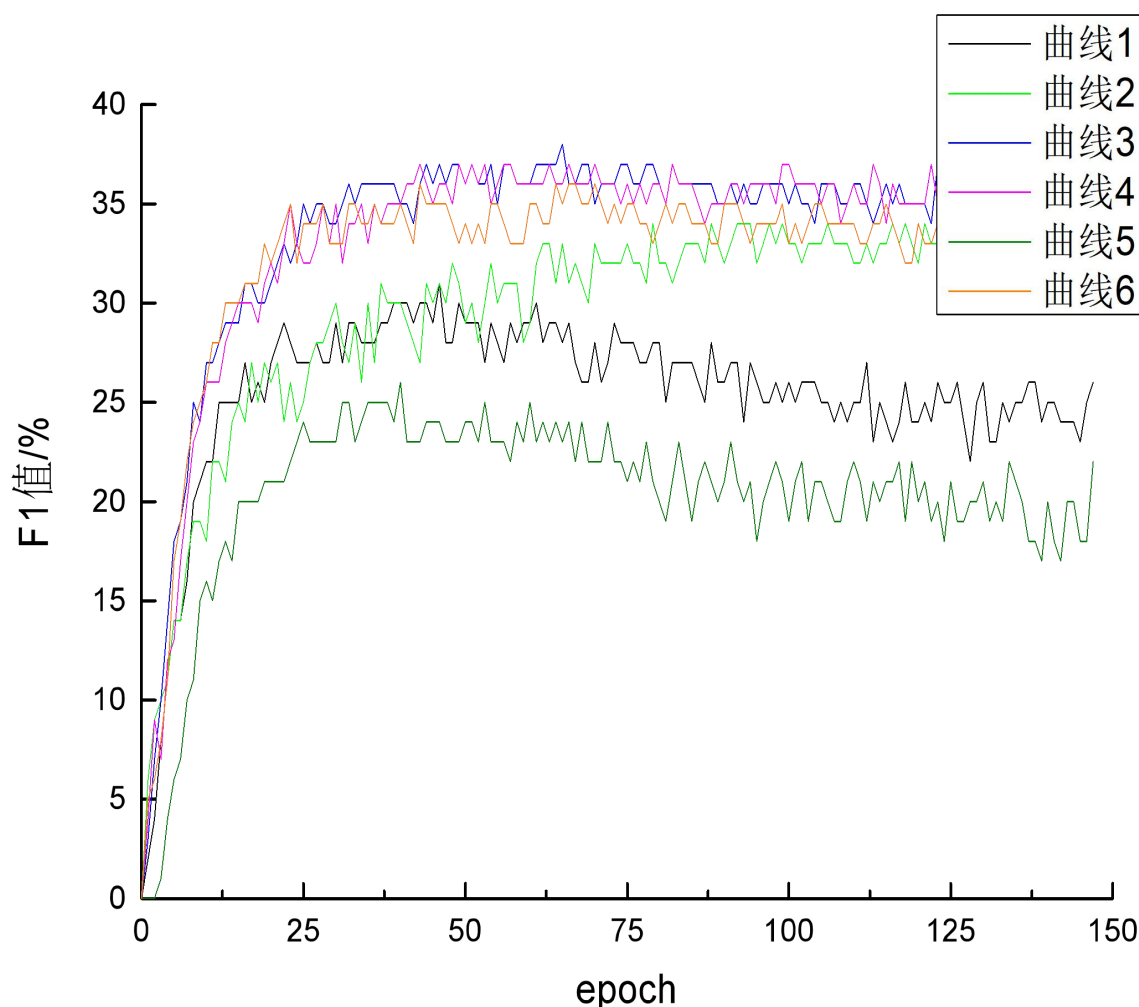


图 4-1 不同参数下的模型性能图

从图 4.1 中可以看出在其它条件相同的情况下, Word2vec 模型采用不同参数训练得到的词向量对标注模型的性能也具有很大影响, 尤其是曲线 1 和曲线 5 中 sg 取值 0 即采用 CBOW 模型进行训练时, 模型的标注性能比其它训练参数的要低很多, 这是因为 CBOW 模型对低频词汇的学习效果相对于 Skip-gram 模型要差一些, 而数据集中低频词汇比较常见。相比于训练模型 sg 参数, 其它参数对模型的标注性能影响要小很多, 通过标注性能对比可以发现曲线 3 对应的标注性能最佳, 所以之后所有的实验均采用曲线 3 对应的参数对 Word2vec 模型进行训练。

4.4 词向量维数对模型性能的影响

通过前文对本文标注模型工作原理的分析可以得出本文模型输出层神经元数目只与 Word2vec 模型输出的词向量维度相关, 与数据集的标注词汇量不再相关。

为了进一步探究 Word2vec 模型输出的词向量维度对模型标注性能的影响，也为了证明本文模型的输出层神经元数目可以进行自由选择，本文对不同 Word2vec 词向量维度下的模型标注性能进行测试。模型在测试时，除了 Word2vec 模型的输出词向量维度改变以外，标注模型的所有参数均相同，为了减少训练时间和避免训练过程中调参的影响，模型损失函数只采用生成对抗损失。图 4.2 与 4.3 为基于 Corel 5K 和 IAPRTC-12 数据集不同词向量维度下的模型标注性能图，在作图时选择 F1 最大时的准确率 P、召回率 R 和 F1 值作为模型在该词向量维度下的最佳标注性能。

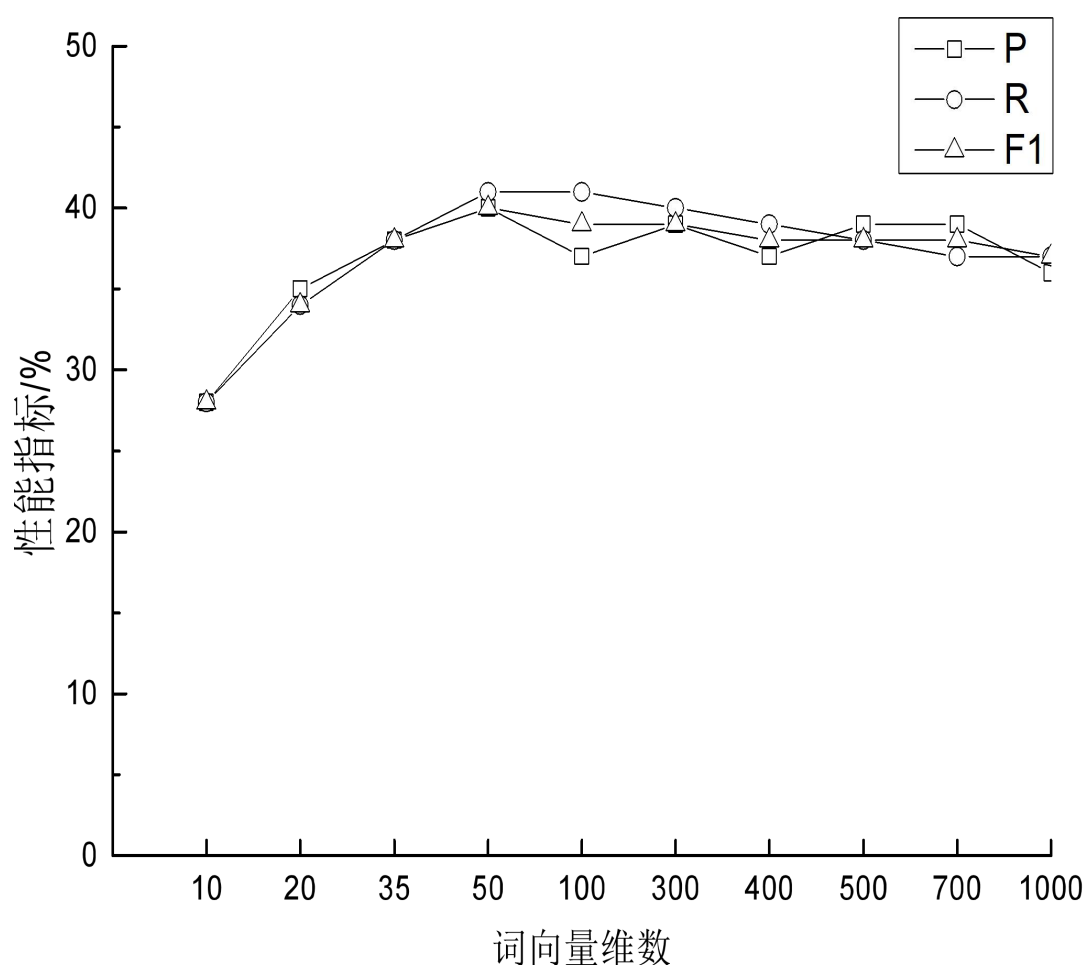


图 4-2 基于Corel 5K数据集不同词向量维数下的模型性能图

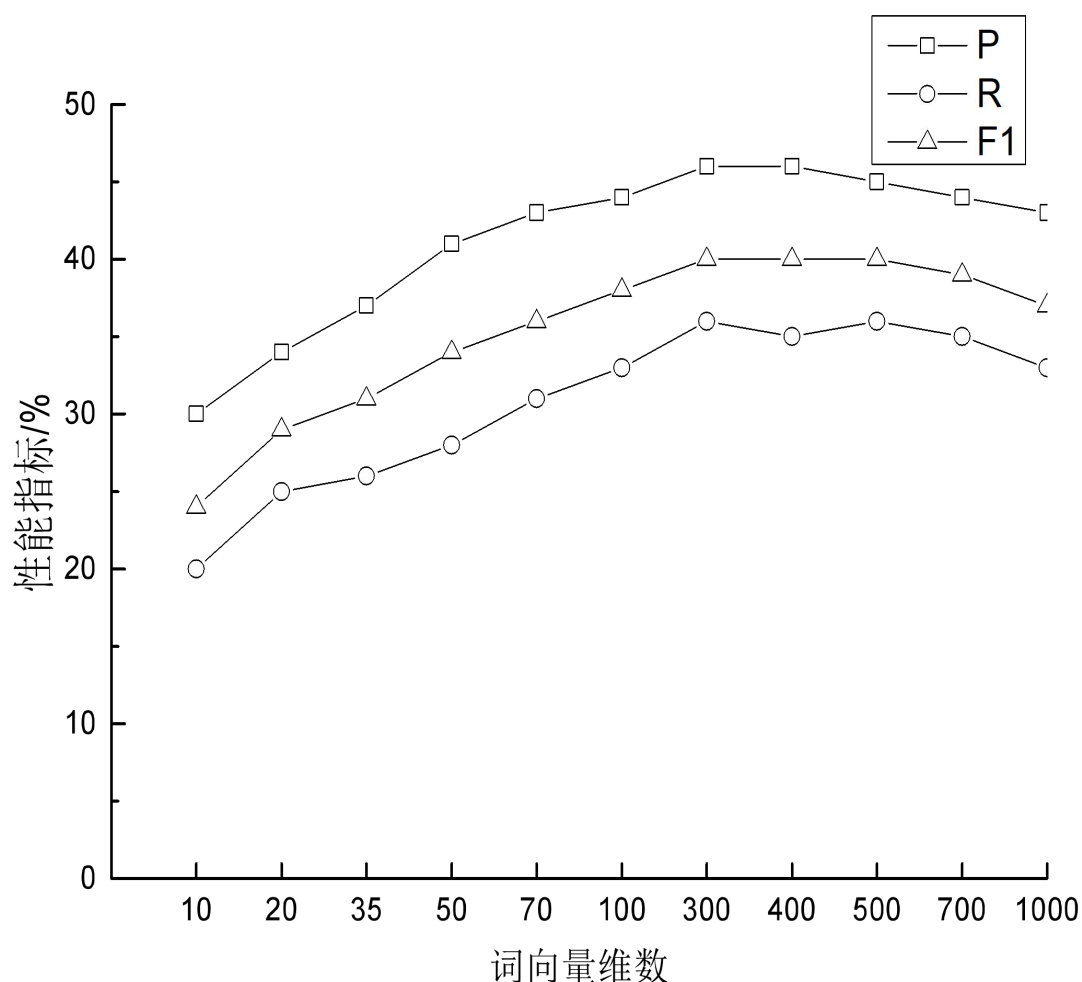


图 4-3 基于IAPRTC-12数据集不同词向量维数下的模型性能图

从图中可以看出：词向量维数在 35 到 1000 维之间时（对于 IAPRTC-12 数据集为 70 到 1000，因为其标注词汇量大于 Corel 5K 数据集），模型的标注性能无论是准确率 P、召回率 R 和 F1 值的差异都十分小，最大差异为 3 个百分点，这些差异可能由深度学习模型的随机性、词向量维数的变化、不同输出层神经元数目与整个模型网络的匹配度等多种原因共同引起。由于差异十分小，并且产生的因数较多，所以可以近似的认为模型词向量维数在 35 到 1000 维之间时，词向量维数的选择对于模型的最终标注性能没有影响。

当词向量维数小于 35（对于 IAPRTC-12 数据集为 70）时，模型的标注性能开始快速下降，其原因为：1.标注模型的整体结构是按照输出层神经元数目为 500 时设计的，当词向量维数太小时导致模型输出层神经元数目过小，与模型其它部分结构不匹配，使得模型整体结构变得不合理，导致模型性能下降 2.由于 Word2vec

模型输出词向量取值范围为 $[-1,1]$ ，因此当词向量维数的太小时，不同标签词汇对应词向量之间的差异变的非常小，加之激活函数对于输出的非线性处理，使得模型不能对不同标签向量进行准确区分，生成器不能准确的生成测试图像对应的词向量，导致模型性能下降。

总而言之，本文标注模型实现了输出层神经元数目与数据集标注词汇量之间的解绑，输出层神经元数目可以在很大范围内自由选择，这将给标注模型带来诸多优势。与大部分输出层神经元数目与数据集词汇量成比例的标注模型相比，对于标注词汇量较大的数据集而言，本文的标注模型可以选择较小维数的词向量，实现对标注词汇的压缩，减小模型输出层神经元(分类器)数目，例如对于 Open Images 数据集标注模型可以选择 1000 或者 2000 个而不是超过 2 万个神经元的输出层，这将大大减小标注模型的参数量，使得模型的训练时间和过拟合问题减小，同时也更容易设计出一个合理的模型结构；对于标注词汇量较小的数据集而言，通过 Word2vec 模型选择输出一个较大维数的词向量，如本文对于 corel 5k 数据集选择 500 维，这样能减小了模型的设计难度。此外，gensim 的 Word2vec 模型可以实现词向量增量训练功能，在保持原词汇表中单词对应词向量不变的基础上，对新加入的单词进行训练，得到新的词向量转换模型。这样，由于新加入的单词不改变之前单词对应的词向量，这使得本文模型面对数据集内容变化时，可以不改变模型结构，只通过对 Word2vec 模型进行增量训练就可以在不影响之前标注模型训练结果的基础上实现对新增标注词汇的学习，增强了模型的实用性。

4.5 不同阈值对图像标注的影响

阈值参数是本文模型的一个重要超参数，对于模型的最终标注性能具有巨大的影响，为了探究阈值参数对模型标注性能的影响和阈值参数的选择方法，本文通过在不同阈值的情形下的同一经过学习的模型进行性能测试，得到阈值对模型标注的具体影响曲线。图 4-4 及图 4-5 为分别在 Corel5K 和 IAPRTC-12 数据集上模型标注的准确率 P、召回率 R、F1 值与阈值参数大小的关系图。

在进行模型测试时，Word2vec 模型的训练参数选择 4.3 节中的最佳性能曲线对应的参数，输出词向量的维数均选择 500，以利于模型网络的设计；模型对于每

幅图像的预测次数大小选择一个 batch_size, 128 次, Word2vec 模型每次对生成的词向量反向求取最接近词向量的 5 个标注作为当次模型的候选标注结果; 对模型多次标注结果进行统计, 根据阈值参数筛选出出现次数大于阈值的标注作为当前阈值参数下模型对图像的最终标注。对于同一训练好的模型, 通过改变阈值参数的取值获取不同阈值下模型的性能, 绘制成以下阈值与模型性能的关系图。

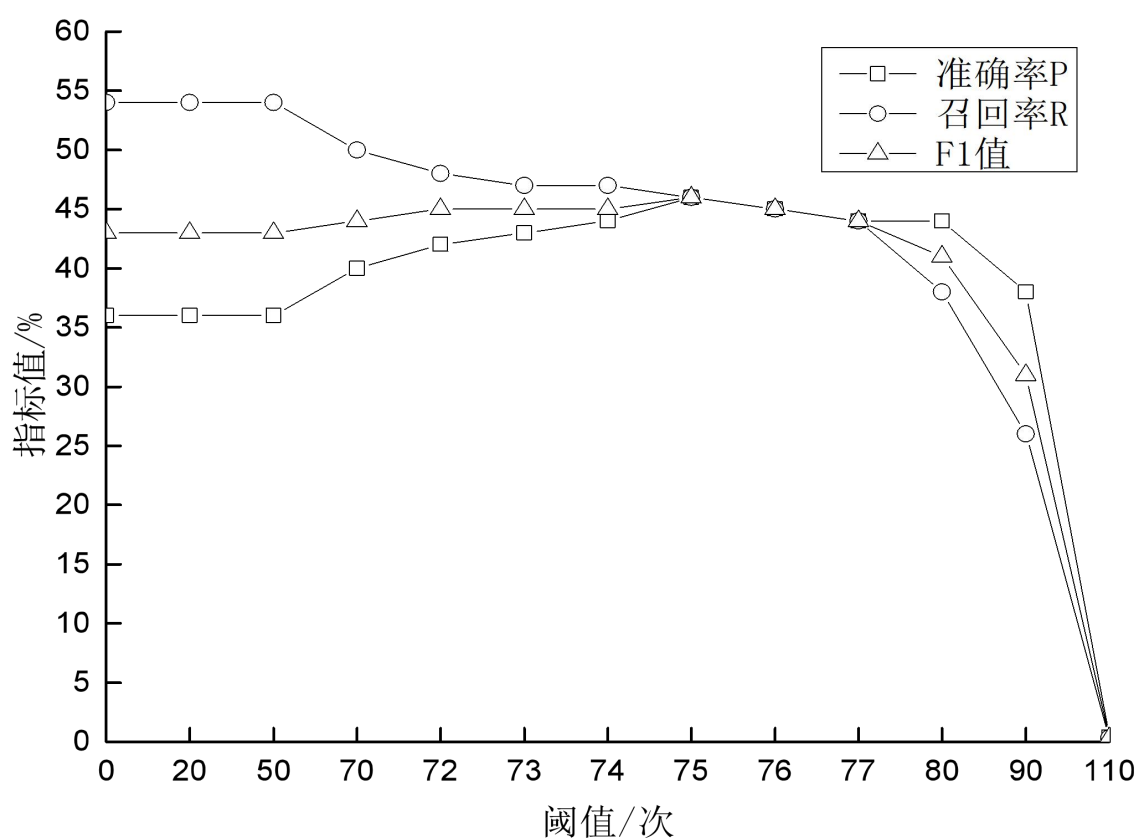


图4-4 Corel 5K数据集下阈值性能影响图

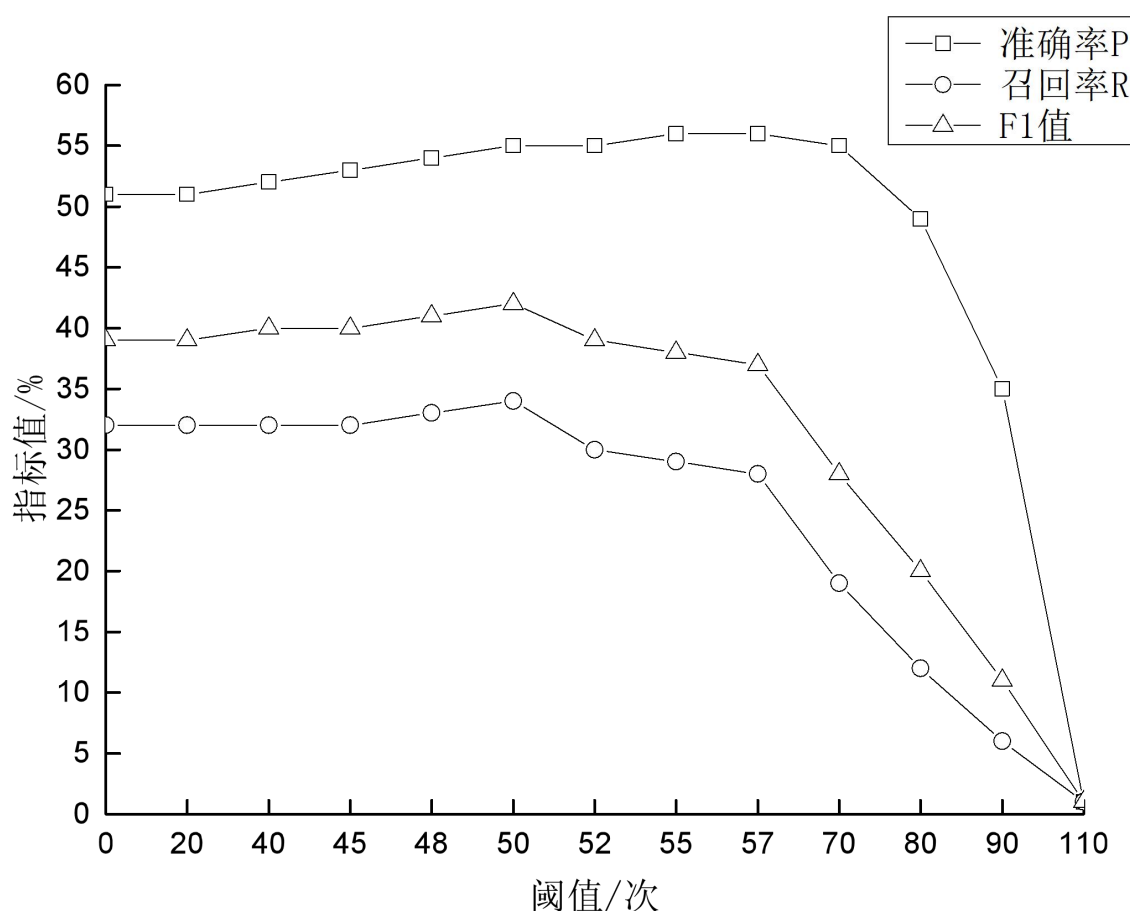


图 4-5 IAPRTC-12数据集下阈值性能影响图

从图 4-4 和图 4-5 可以看出模型性能随阈值变化的大致规律:模型标注的准确率 P 基本随阈值的增长呈现先不变再缓慢上升后快速下降的规律;召回率 R 随阈值的增长先基本不变而后开始下降直到为 0; $F1$ 值的变化规律由准确率 P 和召回率 R 的变化共同决定,基本上表现为随阈值增长先不变然后开始略微上涨最后下降为 0。

阈值对模型性能的影响呈现这种规律的原因大致为:首先,通过训练集对模型的训练,模型可以从数据集中学习到图像特征与标签向量之间的映射关系,使得模型对于新图像具有一定的标注能力。其次,由于标注数据的分布不均匀导致标注结果中,标注正确的标签一般为数据集中在出现频次较高的标签,这些标签在模型标注时出现的次数一般也比较高,同时由于部分标签在数据集中出现频次太少,模型无法有效的学习到这些标签对应的图像特征,导致模型标注时偏向于忽略这些标签,使得模型对测试图像进行标注时,标注对应的出现次数都不会太小。所以,当阈值较小时,因为模型给出的标注对应的出现一般次数大于阈值,

基本没有被阈值过滤掉的标注, 模型标注的准确率 P 和召回率 R 都不变, $F1$ 值不变; 阈值增加到一定值时, 部分标注的出现次数小于阈值, 开始被过滤掉, 又因为标注正确的标签一般出现频次较高, 因此在这部分被过滤掉的标注中, 错误的标注的占比较大, 因此在整体标注中错误标注数目开始减少, 正确标注数目基本不受影响, 导致模型标注的准确率 P 上升, 召回率 R 基本不变, $F1$ 值开始略微上升; 随着阈值继续增加, 正确的标注也开始被过滤, 由于标注正确的标签一般集中在较高的出现频次, 导致在被过滤的标注中正确的标注占比快速增加, 在整体标注结果中, 正确标注被过滤的数目增速远超错误标注被过滤的增速, 最终使得模型标注准确率 P 减小, 召回率 R 都将减小, $F1$ 值随之下降, 直到阈值过大使得所有正确的标注都被阈值完全过滤掉, 模型的标注准确率 P 和召回率 R 都为 0, $F1$ 值也减为 0。

由于阈值对模型的最终标注性能影响巨大, 为了使模型达到最佳性能, 也为了与其它标注模型进行性能对比, 需要对模型的最佳阈值进行确定。由于 $F1$ 值能兼顾准确率 P 和召回率 R 的变化, 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的最佳阈值。由于不同数据集之间存在的差异会导致对于不同数据集模型的最佳阈值也可能不相同, 所以对于 Corel 5K 和 IAPRTC-12 数据集, 在本文模型分别选择 75 和 50 作为模型的最佳阈值。

4.6 不同模型标注性能对比

为了对本文模型的标注性能水平有一个更加明确的定位, 本文将模型的性能指标与多个经典的标注方法进行对比, 这里涉及的对比方法包括传统的机器学习标注方法: RF-opt(Random Forest-Optimize)^[49]、2PKNN、2PKNN-ML(2PKNN-Metric Learning)、SKL-CRM、KSVM-VT^[50]和基于深度学习的图像标注方法 NN-CNN(NearestNeighbor-CNN)^[51]、CNN-R(CNN-Regression)^[52]、ADA(Attribute Discrimination Annotation)^[53]、SNDF(Automatic Image Annotation Combining Semantic Neighbors And Deep Features)^[54]、CNN-MSE、CNN-MLSU。表 3 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 3 模型性能对比表

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	49	37	42	38	44	41
GAN-W	46	46	46	34	55	42

通过表 3 可以看出：

在 Corel 5K 数据集上，本文模型与传统机器学习方法相比准确率 P 和 F1 均为最高，比 RF-opt 方法分别提高 17 和 12 个百分点，召回率与 SKL-CRM、2PKNN-ML 方法取得并列第一，比 RF-opt 方法提高 6 个百分点；与深度学习的标注方法相比，准确率 P 和 F1 也均为最高，分别比 ADA 方法提高 14 和 10 个百分点，召回率仅低于 CNN-MLSU 方法 3 个百分点，但是准确率 P 和 F1 远高于 CNN-MLSU 方法。

在 IAPRTC-12 数据集上，模型也有良好表现，与传统机器学习方法相比准确率 P、召回率 R 和 F1 值均为第一，比 RF-opt 方法分别提高 3 个、11 个和 6 个百分点；与深度学习的标注方法相比，准确率 P 和 F1 也均为最高分别比 ADA 方法提高 13 和 7 个百分点，召回率仅低于 CNN-MLSU 方法 4 个百分点，但是准确率 P 远高于

CNN-MLSU 方法。

综合 GAN-W 模型在 Corel 5K 和 IAPRTC-12 数据集上的性能对比数据可以得出，GAN-W 模型与其它图像标方法相比，准确率 P 和 F1 值表现较好，达到最佳效果，模型召回率 P 虽然略微低于 CNN-MLSU 方法，但是与其他模型相比效果依然不错，并且在整体性能上也优于 CNN-MLSU 方法。所以，在总体上本文 GAN-W 模型性能优于其它模型，在各个指标上都具有明显的提升，可以很好地应用于图像标注领域。

4.7 模型实际标注效果

图 4-6 中给出本文模型的实际标注结果，在模型进行标注时预测次数选择一个 batch_size，128 次，对于 Corel 5K 数据集和 IAPRTC-12 数据集模型选择的阈值分别为 75 和 50，每幅图像选出现次数大于阈值的标注作为该图像的最终标注。

数据集	图像	原始标注	模型标注
Corel 5K		city,sun,water	city, sun, Water, sky
		jet,plan,f-16	f-16, jet, plan
		coral, anemone, ocean, reefs	coral, anemone, ocean, reefs
		tree, horses, mare, foals	horses, tree, mare, foals
		bear, polar, snow, face	bear, polar, snow, tundra






IAPRTC-12		car, desert	car, desert
		child, kid, table	child, girl, desk
		hill	sunset, hill, landscape
		brick, building, child, rock	rock, building, child, brick
		hill, landscape, mountain, rock, woman	hill, woman, mountain, rock, landscape

图 4-6 模型实际标注效果

从模型的实际标注效果图中可以看出：

(1) 与大部分现有标注模型不同，本文模型的对每幅图像给予的标注数目不是一个定值，不同图像可能有不同的标注数目，图像对应的实际标注数目可以做到自适应，这样的标注更加符合实际标注情况。本文模型标注数目自适应的原因：通过对模型的训练，模型可以学到图像特征与标签向量之间的映射关系，模型在每次对图像进行标注时，就会根据被预测图像视觉特征中的某种特征输出一个与之对应的词向量。对于某些图像，由于包含内容比较单一，所以其图像视觉特征只与某些个的标签相对应，所以模型每次标注的向量基本上都接近与该标签向量，使得模型最终标注数目较少；对于包含内容比较复杂的图像，其图像视觉特征可能包含与多个标签向量相对应的图像特征，所以通过模型随机噪声的扰动，

使得这些标签向量中每一个都有概率成为模型标注的输出词向量，通过多次标注预测之后，这些标签中的每个标签出现次数都不会太小，不会被阈值参数给过滤，使得模型可以输出多个标注。此外，数据集也对模型标注数目有影响，由于数据集的限制，导致图像许多内容没有对应的标注，使得某些图像虽然内容复杂但是模型实际给予的标注数目很少。

(2) 模型给出的某些标注虽然与原数据集标注不符合，但是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 sky 与 sun、water）在数据集中共现频率较高，使得这些标注在使用 Word2vec 模型进行向量化时，他们对应的词向量之间的距离很近，所以在通过 Word2vec 模型反向获取生成词向量时，这些与原标注不符合的标注依然有较大概率被 Word2vec 模型一起输出，并且标注与生成词向量匹配的概率也很大，导致最终出现次数依然很大，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存在较深的关联，如 sky 与 sun，所以在新的测试图像中，这些标注对应的图像特征依然有较大概率一起出现，只是在测试数据集中由于标注人员的差异或者不是主要图像特征而被忽略掉。例如上图第一幅测试图像中，sky 的特征虽然在图中有体现但是并不是该测试图像的主要内容，导致 sky 并不在原始测试集的标注中，但是 sky 标签在数据集中多与 sun、water 标签一起出现，加之图像中也隐含了 sky 的对应的图像特征，所以模型认为 sky 符合该测试图像的语义，最终被输出作为该测试图像的标注之一。

4.8 本章小结

本章通过具体的实验对本文模型的有效性进行验证。首先对数据集、标注性能评价方法进行简单介绍。其后通过多方面的实验对模型进行测试：1、通过在不同 Word2vec 模型参数条件下的标注实验，说明 Word2vec 模型参数对模型标注性能的影响 2、通过不同 Word2vec 模型词向量维数对标注性能影响的实验，证明本文模型解决了输出层神经元数目与标注词汇量绑定的问题，本文模型输出层神经元数目可以在一个较广的范围自由选择 3、通过在不同阈值下的模型性能的测试对比，分析阈值超参数对模型标注性能的影响，给出阈值超参数的选取方法 4、通过将模型与多个传统机器学习标注模型和深度学习标注模型进行性能对比，得出本文

模型相比于其它对比模型在性能上都有一定的提高，验证了本文模型的有效性 5、通过模型的实际标注结果实验证明模型标注数目自适应和错误标注仍可能与测试图像相关联的优点，并对其原因进行具体分析。本章通过具体的实验证明了本文模型能在解决输出层神经元数目与标注词汇量成比例的问题的同时相比于其他标注模型还能够进一步提升标注性能，在实际标注中同样具有优势，说明本文模型的有效性和实用性。

5 总结与展望

近年来，图像数据的快速增长使得图像自动标注成为了一个热门研究方向。本文首先介绍了图像标注的研究背景和国内外的研究情况，从中分析出当前基于深度学习和部分基于传统机器学习的图像标注模型具有输出层神经元数目与标注词汇量成比例的问题，并且深入分析其带来的不良后果。

其后，本文对卷积神经网络、生成式对抗网络、Word2vec 模型、迁移学习等模型/方法的原理进行介绍，并在这些技术的基础上设计实现基于生成式对抗网络和词向量模型 Word2vec 的一种新标注模型，用于解决输出层神经元数目与标注词汇量成比例的问题。标注模型利用 Word2vec 模型将标注词汇映射成一个维度可自由选择的多维词向量，再利用生成式对抗网络生成结果的多样性，将大多数模型一次性输出测试图像所有标注的结构改为每次输出一个与词向量维度相同的候选标注，使得模型输出层神经元数目只与词向量的维度相关，而词向量的维度可由开发者自由选择，与数据集标注词汇量无关，解决之前分析出的模型缺陷。

之后，本文介绍了模型的数据预处理方式，给出了模型的生成器与判别器的结构，在生成式对抗网络损失的基础上引入 L2 损失、感知损失，通过加权的方式构成模型的最终损失来提高模型的标注性能，也对模型的训练和测试流程算法进行简单说明。

最后，在 Corel 5K 和 IAPRTC-12 数据集上对本文的标注模型进行实验。通过 5 个不同方面的实验，证明了模型能够解决了本文分析出的模型输出层问题，输出层神经元数目可自由选取，与此同时，和其它经典标注模型相比在各个性能指标上都有较大提升，模型的实际标注结果也有标注数目自适应等优点。

本文模型虽然解决了输出层神经元数目与标注词汇量成比例的问题，在此同时本文模型相较于其它经典模型标注性能也有较大提升，然而由于实验条件有限，以及研究水平的不足，使得本文的标注模型还有一些值得改进和进一步研究的地方：

- 1、由于实验条件和时间的限制，模型许多参数未经过细致调整，导致模型未被调整到最佳状态，模型的性能仍可以进一步提高；另外，本文实验只在 Corel 5K

和 IAPRTC-12 这两个常用数据集上进行,未来可以在如 NUS-WIDE 数据集、Open Images 数据集等更多数据集上对模型进行进一步实验。

2、本文模型中的生成器和判别器采用简单的全连接层构成,结构比较简单,未来可以对生成器和判别器的结构进行进一步优化,提高模型的标注性能。

3、当前对于生成的词向量缺乏一个较好的直接评判标准,在 Word2vec 模型的学习过程中也没有一个明确的指标来指导参数的选择与调整,未来可以对 Word2vec 模型进一步研究,给出模型词向量质量的评判方法来指导 Word2vec 模型的训练,提升模型的标注性能。

致谢

时光飞逝，研究生的三年学习生涯已经进入尾声，在三年的研究生生活中，老师、同学、家人都给予了许多帮助，使我在学习和生活中都得到了锻炼和成长。在这论文完成之时，向他们表达我的感谢之情。

首先，我要感谢实验室的各位老师，特别是我的指导老师刘卫忠老师。刘老师在我三年的学习生活中，在学术研究上给予我许多悉心指导，在实践操作上也给予我许多机会，参与了多个项目研发，提高了我的工程能力和实践经验。对于我的毕业论文，无论是论文的选题，还是中期研究实验，亦或是最后论文的写作，刘老师都给予细心的指导，给出了详细的参考意见。在此向我的导师表达深切的感谢！

其次，也要对我的实验室同学和朋友表达感谢，在三年的生活中，互相帮助，一起度过了快乐研究生生活；在学术上，遇到的问题都会一起进行探讨，不仅让基础知识的学习变得简单，而且在毕业论文的各个实验中遇到的许多问题也都在探讨中被轻松解决掉。

最后，要对我的家人表达特别感谢。他们在我的整个人生中一直都给予无私的帮助与关心。在我感到迷茫时，在我遇到困难时，都是他们一直陪伴着我，给予我安慰和信心，支持我一直走下去。

参考文献

- [1] 李敬伟.基于多标记学习的图像标注算法研究与实现[D].北京交通大学,2017.
- [2] Mary Meeker. Internet Trends 2016-Code Conference[R].Los angels: kleiner perkins caufield & byers. 2016.
- [3] 宋光慧. 基于迁移学习与深度卷积特征的图像标注方法研究[D]. 2017.
- [4] Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, et al. Multi-label classification via calibrated label ranking [J]. Machine Learning, 2008, 73 (2):133-153.
- [5] Elisseeff A E , Weston J . A kernel method for multi-labelled classification[C]// Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. MIT Press, 2001.
- [6] Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007,40(7):2038-2048
- [7] 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.
- [8] 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.
- [9] 李志欣,郑永哲, 张灿龙,等. 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.
- [10] 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.
- [11] 李志欣, 施智平, 李志清, 等. 融合语义主题的图像自动标注[J]. 软件学报,
2011, 22(4):801-812.
- [12] Lavrenko V, Manmatha R, Jeon J. A model for learning the semantics of pictures
[J]. Nips, 2004: 553-560.

- [13] 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.
- [14] Moran S, Lavrenko V. A sparse kernel relevance model for automatic image annotation [J].Journal of Multimedia Information Retrieval, 2014, 3(4):209-229.
- [15] Makadia A, Pavlovic V, Kumar S. Baselines for image annotation[J]. International Journal of Computer Vision, 2010, 90(1):88-105.
- [16] 李志欣, 施智平, 张灿龙, 等. 混合生成式和判别式模型的图像自动标注[J]. 中国图像图形学报, 2018, 20(5):687-699.
- [17] 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.
- [18] 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.
- [19] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7):2038-2048.
- [20] Gao Y L, Fan J P, Xue X Y, et al. Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers // Proceedings of the 14th ACM International Conference on Multimedia. New York, USA: ACM, 2006:901-910.
- [21] 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.
- [22] Carneiro G, Chan A B, Moreno P J, et al. Supervised learning of semantic classes for image annotation and retrieval [J]. IEEE Transactions on Pattern Analysis and

- Machine Intelligence, 2007, 29(3):394-410.
- [23] Vallet A, Sakamoto H. A multi-label convolutional neural network for automatic image annotation [J]. Journal of Information Processing, 2015, 23(6), 767-775.
- [24] 汪鹏, 张奥帆, 王利琴, 等. 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.
- [25] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554.
- [26] 尹宝才, 王文通, 王立春. 深度学习研究综述[J]. 北京工业大学学报, 2015 (1):48-59.
- [27] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20(3):273-297.
- [28] Schapire R E. The strength of weak Learnability [J]. Machine Learning, 1990, 5(2):197-227.
- [29] Cox D R. The regression analysis of binary sequences[J]. Journal of the Royal Statistical Society, 1958, 20(2):215-242.
- [30] Lecun Y, Boser B, Denker J S, et al. Handwritten digit recognition with a back-propagation network[C] //Advances in Neural Information Processing Systems. CA: Morgan Kaufmann Publishers, 1990:396-404.
- [31] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [32] Salakhutdinov R, Hinton G E. Deep Boltzmann machines [C] //JMLR Workshop and Conference Proceedings Volume 5: AISTATS 2009. Brookline, MA: Microtome Publishing, 2009: 448-455.
- [33] 刘建伟, 刘媛, 罗雄麟. 玻尔兹曼机研究进展[J]. 计算机研究与发展, 2014, 51(1): 1-16.
- [34] Cho K, Van M B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. CoRR, 2014: abs/1406.1078.

- [35] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2012, 25(2):2012.
- [36] 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.
- [37] 陈锐, 乔沁, 宋志坚. 生成对抗网络在医学图像处理中的应用[J]. 生命科学仪器, 2018, 16(Z1):73-82+93.
- [38] 王坤峰, 苟超, 段艳杰, 等. 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.
- [39] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784,2014.
- [40] Arjovsky M, Chintala S, Bottou, Léon. Wasserstein GAN [J]. arXiv preprint arXiv:1701.07875, 2017.
- [41] 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.
- [42] Mikolov T.Word2vec [EB/OL]. [2014-04-15]. <http://code.google.com/p/Word2vec>
- [43] Kenter T, Borisov A, Rijke MD. Siamese CBOW: Optimizing word embedding for sentence representations[C]//Meeting of the Association for Computational Linguistics, 2016:941-951.
- [44] Shazeer N, Pelemans J, Chelba C. Skip-gram language modeling using sparse non-negative matrix probability estimation [J]. Computer Science, 2015.
- [45] Pan S J, Yang Q. A survey on transfer learning [J]. IEEE Transactions on Knowledge &Data Engineering, 2010, 22(10): 1345-1359.
- [46] 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.
- [47] P Duygulu, K Barnard, J F G de Freitas, et al. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary [J].Lecture Notes in Computer Science,2006 (2353):349-354.
- [48] Grubinger, M., Clough, P. and Muller, H. The IAPR Benchmark: A new evaluation resource for visual information systems. International Conference on Language Resources and Evaluation, 2006,3-23.
- [49] 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
- [50] 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.
- [51] 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.
- [52] 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.
- [53] 周铭柯, 柯逍, 杜明智. 基于数据均衡的增进式深度自动图像标注[J]. 软件学报, 2017, 28(7):1862-1880.
- [54] 柯逍, 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.

攻读硕士学位期间发表学术论文情况

[1] 税留成, 刘卫忠, 冯卓明. 基于生成式对抗网络的图像自动标注[J/OL]. 计算机应用 :1-6[2019-03-26]. <http://kns.cnki.net/kcms/detail/51.1307.TP.20190227.1237.002.html>