

学校代码： 10246

学 号： 072021109

復旦大學

硕 士 学 位 论 文

基于语义上下文建模的图像语义自动标注研究

院 系： 计算机科学技术学院

专 业： 计算机软件与理论

姓 名： 向 宇

指 导 教 师： 周向东 副教授

完 成 日 期： 2010 年 5 月 1 日

指导小组成员名单

周向东 副教授

计算科学技术学院

复旦大学

施伯乐 教授

计算科学技术学院

复旦大学

目录

摘要.....	v
Abstract.....	vii
第一章 绪论	1
1.1. 引言	1
1.2. 本文工作	3
1.3. 本文组织结构	5
第二章 相关工作及研究背景	6
2.1. 图像语义自动标注	6
2.2. 语义上下文建模	7
2.3. 马尔科夫随机场	8
第三章 多马尔科夫随机场上下文相关模型	10
3.1. 概念图	11
3.2. 基于生成模型的势函数设计	11
3.3. 正则化最大伪似然参数估计	12
3.4. 模型推理	14
3.5. MMRF 图像语义自动标注算法	15
3.5.1. 训练集的构造.....	15
3.5.2. 标注算法.....	16
3.6. 实验	16
3.6.1. 实验数据集.....	16
3.6.2. 评价度量.....	17
3.6.3. 在 Corel 数据集上的对比.....	17
3.6.4. 在 TRECVID-2005 数据集上的对比	20
3.7. 本章小结	21
第四章 最大边缘条件随机场上下文相关模型	22
4.1. 条件随机场	23
4.2. 概念图	23
4.3. 势函数设计	24
4.4. 最大边缘参数估计	25
4.4.1. 拆分的 Hinge 损失.....	25
4.4.2. 有偏向的正则化.....	27

4.4.3. 参数估计框架.....	27
4.4.4. 利用上下文核函数求解最优化问题的算法.....	27
4.4.5. 核函数的构造.....	28
4.5. 模型推理	29
4.6. 实验	30
4.6.1. 实验数据集.....	30
4.6.2. 特征提取.....	30
4.6.3. 评价度量.....	31
4.6.4. 语义上下文建模评价.....	31
4.6.5. 在 Corel 数据集上的对比.....	32
4.6.6. 在 TRECVID-2005 数据集上的对比	33
4.7. 本章小结	35
第五章 结束语	36
5.1. 本文贡献.....	36
5.1.1. 马尔科夫随机场标注框架.....	36
5.1.2. 多马尔科夫随机场上下文相关模型.....	36
5.1.3. 最大边缘条件随机场上下文相关模型.....	37
5.2. 将来工作	37
附录.....	38
1. 命题 4.1 的证明.....	38
参考文献.....	40
攻读学位期间作者的研究成果.....	45
1. 参与科研项目	45
2. 已发表和录用论文.....	45
致谢.....	46

摘要

由于图像语义自动标注 (Automatic Image Annotation, AIA) 在基于关键词的图像和视频的检索与浏览上具有巨大的应用前景, AIA 在近年来受到了人们的广泛关注。解决 AIA 问题的瓶颈在于图像底层的视觉特征与高层的语义概念之间存在“语义鸿沟” (Semantic Gap), 即图像视觉特征相似并不能保证图像语义一致。为了跨越这条“语义鸿沟”, 研究者们基于生成模型和判别模型提出了多种图像语义自动标注的方法。此外, 语义概念之间的相互关系已经被应用于图像语义自动标注, 并且取得了令人鼓舞的结果。通过对语义上下文建模, 生成模型和判别模型的性能都得到了改进。

本研究工作提出了一个马尔科夫随机场 (Markov Random Field, MRF) 标注框架用于对图像语义自动标注中的语义上下文建模。与先前视觉识别工作中对图像像素或图像区域空间位置关系建模的 MRF 不同, 我们提出的 MRF 是在语义概念上构造, 用于对语义概念之间的相互关系建模。具体来讲, MRF 中的点表示语义概念, 而边表示语义概念之间的相关性。每个点上有一个二值标签来表示相应的语义概念在给定的图像中出现或不出现。

在 MRF 标注框架下, 我们提出了一种新颖的多马尔科夫随机场 (Multiple Markov Random Field, MMRF) 上下文相关模型对语义上下文建模。MMRF 通过构造语义层的 MRF 模型来改进 AIA 中传统生成模型的标注结果。具体来讲, 我们基于生成模型估计的图像视觉特征与语义概念共同出现的联合概率, 设计了 MRF 新颖的势函数。为了准确地捕获不同语义概念的语义, 我们为每一个语义概念构造自身的 MRF。此外, 我们高效地解决了 MMRF 的参数估计和模型推理问题。

为了进一步发掘语义上下文相关模型的能力, 我们在 MRF 标注框架下提出了一种新颖的判别条件随机场模型对语义上下文建模, 称之为最大边缘条件随机场 (Maximal Margin Conditional Random Field, MMCRF) 上下文相关模型。MMCRF 能够同时从语义层次与视觉层次上对语义相关性建模。具体来讲, 我们基于线性判别模型设计了 MMCRF 的势函数, 并提出了拆分的 Hinge 损失在最大边缘框架下估计 MMCRF 的参数。模型的训练转化为采用我们推导出的上下文核函数求解一系列独立的二次规划问题。

我们在公用的标注数据集: Corel 图像数据集和 TRECVID-2005 视频数据集上进行了实验来评估 MMRF 和 MMCRF 的标注性能。实验结果表明, 与当前最先进的标注方法相比, 我们的模型能够显著地改进标注性能。特别是 MMRF 在 Corel 数据集 263 个关键词上的平均查全率和平均查准率分别达到了 0.36 和 0.31, 至今仍然是 Corel 数据集上一个很有竞争力的结果。

关键词：图像语义自动标注，语义上下文建模，马尔科夫随机场，条件随机场，生成模型，判别模型，最大边缘

中图法分类号：TP311

Abstract

Automatic Image Annotation (AIA) has attracted increasing attentions in recent years due to its potential in many interesting applications, such as keyword based image and video retrieval and browsing. However, a major bottleneck of AIA is the so-called semantic gap problem between visual perception and high-level semantics. To deal with this challenge, various AIA methods, mostly based on generative models and discriminative models, have been proposed in the current literature. Besides, the relationships between semantic concepts have been utilized in AIA and bring promising results. Semantic context modeling has been integrated with both generative models and discriminative models to leverage the learning power of AIA methods.

This thesis presents a novel Markov Random Field (MRF) annotation framework for semantic context modeling in AIA. Different for the previous MRFs in vision recognition which model the spatial relationships between image pixels or regions, our MRF is built over semantic concepts to model the interactions between them. Specifically, the sites in our MRF correspond to semantic concepts and the edges represent the correlations between concepts. A binary label is associated with each site to indicate the presence or absence of the corresponding concept in an image.

We propose a novel Multiple Markov Random Field (MMRF) contextual model for semantic context modeling in our MRF annotation framework. MMRF builds semantic level MRFs to refine the annotation results of traditional generative models in AIA. Specifically, we propose new potential functions based on joint probabilities of image visual features and semantic concepts estimated by some generative model in AIA. We build one MRF for each semantic concept to capture different semantics among them. Besides, we efficiently solve the parameter estimation and model inference problems of MMRF.

We propose a novel discriminative Conditional Random Field model for semantic context modeling in our MRF annotation framework, called Maximal Margin Conditional Random Field (MMCRF) contextual model. Our model captures the interactions between semantic concepts from both semantic level and visual level in an integrated manner. Specifically, the potential functions are designed based on linear discriminative models, which enable us to propose a novel decoupled Hinge loss function for maximal margin parameter estimation. We train the model by solving a set of independent quadratic programming problems with our derived contextual kernel.

Extensive experiments have been conducted on commonly used benchmarks: Corel and TRECVID-2005 data sets to evaluate the annotation performance of MMRF and MMCRF. The experimental results show that compared with the state-of-the-art methods in AIA, our methods achieved significant improvement on annotation performance. Especially, MMRF achieved 0.36 and 0.31 in average recall and average precision respectively on 263 keywords in Corel data set. This remains a very competitive performance on Corel data set.

Keywords: Automatic image annotation, semantic context modeling, markov random field, conditional random field, generative model, discriminative model, maximal margin

Chinese Library Classification Code: TP311

第一章 绪论

1.1. 引言

随着互联网上数字图像和视频数据的不断增加，如何从这些大量的数据中有效地检索出人们所需要的数据成为一个亟待解决的问题。由于人们习惯采用语义概念来描述所需要的图像或视频，最初的解决方案是首先通过人工对图像或视频标注上相应的语义概念，再通过文本检索系统根据语义概念检索图像或视频。人工标注的缺陷在于需要花费大量的人力，对于大量的数据来说是不可行的。因此，研究人员对图像语义自动标注（Automatic Image Annotation, AIA）问题开展了研究，希望通过计算机算法来实现图像语义的自动标注。

图像语义自动标注根据图像的视觉特征，将图像与相应的语义概念（关键词）联系起来。解决此问题的瓶颈在于图像的底层视觉特征与高层语义概念之间存在一条“语义鸿沟”（Semantic Gap）。一方面，图像视觉特征相似并不能保证它们在语义概念上一致；另一方面，同一个语义概念对应的图像在视觉特征上具有多样性。为了跨越这条“语义鸿沟”，研究者们提出了多种图像语义自动标注的方法。总体上讲，这些方法可分为基于生成模型的方法和基于判别模型的方法两大类。生成模型通过估计图像与语义概念（关键词）共同出现的联合概率分布，把联合概率高的语义概念作为图像的标注。采用生成模型来解决图像语义自动标注的一个开创性工作交叉媒体相关模型（Cross-Media Relevance Model, CMRM）[1]。CMRM 通过对图像区域聚类采用一组离散的簇标号来表示一幅图像，可以看作图像是由视觉关键词组成。因此，CMRM 通过统计视觉关键词和语义关键词的出现频率来估计图像的生成概率。Lavrenko 等[2]提出了连续空间相关模型（Continuous-space Relevance Model, CRM）分别采用非参数估计和多项式分布来估计图像生成区域和语义关键词的概率。Feng 等[3]提出了多伯努利相关模型（Multiple Bernoulli Relevance Model, MBRM）采用多伯努利分布取代多项式分类来估计图像生成语义关键词的概率。判别模型把每个语义概念看作一个类，采用最大后验估计以及其他分类技术来解决图像语义自动标注问题。早期的工作关注于从图像中区分特定的语义概念，例如区分室内场景和室外场景[4]，区分城市景色和山水风光[5]。这些工作把图像自动标注看作二分类问题。然而我们在图像自动标注中常常面对的是多个语义概念，把图像自动标注看作多分类问题更为确切。Cusano 等[6]利用多分类支持向量机（Support Vector Machine, SVM）把图像区域分类到预先定义的多个类。Carneiro 等[7]提出了有监督多分类标注（Supervised Multi-class Labeling, SML）模型，采用双层的高斯混合模型来估计类密度。近些年来，除了利用图像的视觉特征



图 1.1 Corel 数据集中的四幅图像以及相应的人工标注的语义概念

外，生成模型与判别模型都开始利用语义概念之间的相关性来提高图像自动标注的性能。

语义概念常常共同出现在同一幅图像中，例如语义概念“bird”和“tree”，“car”和“track”经常一起出现。图 1.1 给出了从 Corel 数据集[8]中取出的四幅示例图像和相应的人工标注的语义概念。直观上讲，如果我们已知一幅图像已经被标注了“bird”或“car”，那么这幅图像被标注为“tree”或“track”的概率就会比较大。此外，如果两个语义概念从不共同出现在同一幅图像中，利用它们之间的相关性能够减少错误的标注。我们把在同一幅图像中共同出现的互相关联的语义概念称作语义上下文（Semantic Context）。生成模型和判别模型都通过对语义上下文建模来改进图像语义自动标注的性能。[9]和[10]基于相关模型，在估计图像与语义概念的联合概率时引入语义概念之间的相关性。Rasiwasia 和 Vasconcelos[11]在生成模型上采用混合 Dirichlet 分布对语义上下文建模。Qi 等[12]提出了相关联多标签（Correlative Multi-Label, CML）标注模型，把图像语义自动标注看作是多标签分类问题。CML 将图像同时分为多个类，并在分类过程中利用语义概念之间的相关性。

现有的语义上下文模型[9, 10, 11, 12]虽然在图像标注性能上取得了改进，但也存在自身的问题。基于相关模型的方法[9, 10]利用训练数据集或外部的统计信息来计算语义概念之间的相关性，缺乏适当的参数设置来对语义上下文建模，导致模型没有足够的能力来利用语义上下文。[11]是一个分层的标注模型。底层通过生成模型将图像视觉特征与语义概念联系起来，高层采用混合 Dirichlet 分布在底层生成模型的基础上对语义上下文建模。模型的分层降低了模型的效率。此外，图像标注的性能受限于底层的生成模型。CML[12]采用多标签分类技术来处理图像语义自动标注，能够在标注过程中利用语义概念之间的相关性。但是 CML 忽略了语义相关性与图像视觉特征之间的联系。另外，CML 采用结构化 SVM（Structural SVM）[13]来求解多标签分类问题。

训练结构化 SVM 需要求解约束数目为语义概念数目指数倍的多项式规划 (Quadratic Programming, QP) 问题, 导致 CML 对于语义空间不具有可延展性。

总之, 由于“语义鸿沟”的存在, 研究者在加强对图像视觉特征利用的同时, 开始利用语义概念之间的相互关系来改进语义自动标注的性能。图像语义自动标注中的语义上下文建模在近年来成为一个热门的研究课题。现有的语义上下文模型证明了利用语义概念之间的相关性能提高标注性能。但不同的模型对于语义上下文的表示方法不同, 对语义上下文的利用率也不同。如何高效地表示、利用语义上下文仍然是语义上下文建模的核心问题。

1.2. 本文工作

在本文中, 我们基于马尔科夫随机场 (Markov Random Field, MRF) [14], 提出了一个新颖的对图像语义自动标注中语义上下文建模的框架。在该框架下, 我们提出了两个新颖的图像语义自动标注上下文相关模型: 一个是多马尔科夫随机场 (Multiple Markov Random Field, MMRF) 上下文相关模型[15], 另一个是最大边缘条件随机场 (Maximal Margin Conditional Random Field, MMCRF) 上下文相关模型[16]。我们在本节中对所提出的 MRF 语义上下文建模框架进行描述, 随后在第三章和第四章中分别介绍 MMRF 和 MMCRF 上下文相关模型。

马尔科夫随机场 (MRF) 对随机变量以及它们之间的相关关系建模。一组随机变量 $\mathbf{y} = \{y_1, y_2, \dots, y_m\}$, $\mathbf{y} \in Y$ 被称作是点集 $S = \{1, 2, \dots, m\}$ 上, 关于邻域系统 $N = \{N_i | i \in S\}$ 的马尔科夫随机场, 其中 N_i 是与点 i 相邻的点, 当且仅当满足下面两个条件:

$$(1) P(\mathbf{y}) > 0, \forall \mathbf{y} \in Y,$$

$$(2) P(y_i | y_{S-\{i\}}) = P(y_i | y_{N_i}), \forall i \in S,$$

其中 $y_A = \{y_i | i \in A\}$, A 表示任意的点的集合。条件(2)说明一个随机变量仅仅和与它相邻的随机变量相关, 这个性质被称为马尔科夫性 (Markovianity)。Hammersley-Clifford 定理[17]指出, 每一个 MRF 都满足下面的概率密度分布:

$$P(\mathbf{y}) = Z^{-1} \times e^{-U(\mathbf{y})}, \quad (1.1)$$

其中

$$Z = \sum_{\mathbf{y} \in Y} e^{-U(\mathbf{y})} \quad (1.2)$$

是一个正规化的常数, 被称为划分函数 (partition function), $U(\mathbf{y})$ 是能量函数 (energy function), 它是所有集簇 (clique) $c \in C$ 上势函数 (potential function) $V_c(\mathbf{y})$ 的和:

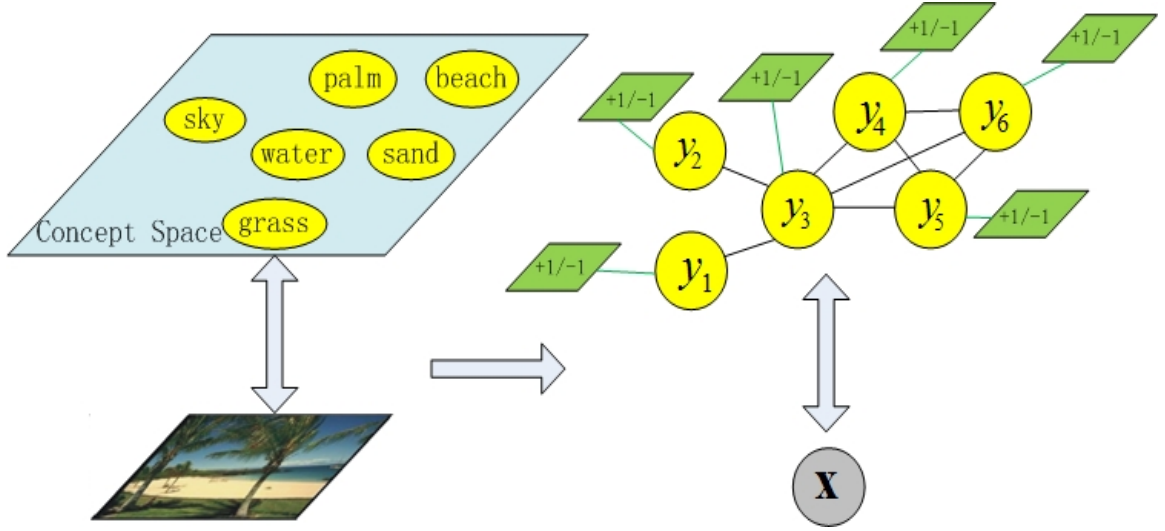


图 1.2 语义上下文建模的图像语义自动标注框架

$$U(\mathbf{y}) = \sum_{c \in C} V_c(\mathbf{y})。$$

当仅考虑点集大小不超过 2 的集簇时，能量函数可简化为：

$$U(\mathbf{y}) = \sum_{i \in S} V_1(y_i) + \sum_{i \in S} \sum_{j \in N_i} V_2(y_i, y_j), \quad (1.3)$$

其中 $V_1(y_i)$ 被称为点势函数(site potential), $V_2(y_i, y_j)$ 被称为边势函数(edge potential)。势函数的设计是构造 MRF 的关键，它使得 MRF 能够适用于不同的应用场景。

在我们提出的马尔科夫随机场标注框架中，我们利用 MRF 来对语义概念及它们之间的相互关系建模。给定一幅输入图像，我们用 $\mathbf{x} \in X$ 表示该图像的视觉特征。图像语义自动标注的任务是根据输入 \mathbf{x} 来预测与之相关的语义概念。语义上下文建模需要在标注过程中利用语义概念之间的相关性。在我们的框架中，MRF 中的每一个点对应一个语义概念，每一条连接两个点的边表示相应的两个语义概念相关。点 i 上的随机变量 $y_i \in \{+1, -1\}$ 表示相应的语义概念在给定的图像中出现或不出现。我们把 y_i 的取值称作点 i 的标签(label)。图像语义自动标注转化为根据 \mathbf{x} 来确定 MRF 点集 S 的标签组合(configuration)，把标签为 +1 的语义概念作为图像的标注。图 1.2 描绘了我们语义上下文建模的图像语义自动标注框架。利用这个框架实现图像语义自动标注需要完成以下四步工作：

1. 构造 MRF 的图结构
2. 设计点势函数和边势函数
3. 估计 MRF 的参数
4. 在 MRF 上进行推理(inference),

其中第 3 步参数估计求出设计势函数时引入的未知参数，第 4 步推理求出 MRF 的标签组合。在这篇论文中，我们分别从生成模型和判别模型出发，基于上述框架提出了

下面两个语义上下文相关模型：

- 多马尔科夫随机场 (MMRF) 模型[15]：MMRF 在底层生成模型的基础上，利用 MRF 为每一个语义概念分别构造高层的语义模型。MMRF 引入了适当的参数设置对语义概念之间的相关性建模，并能够针对不同语义概念高效地学习不同的参数，大大提高了生成模型对语义上下文的利用。
- 最大边缘条件随机场 (MMCRF) 模型[16]：MMCRF 把输入图像看作一个整体，利用条件随机场 (Conditional Random Field, CRF)[18] 对语义上下文建模。MMCRF 采用线性判别模型来定义势函数，其中边势函数可以看作是与输入数据相关的平滑函数，使得 MMCRF 同时从语义层次与视觉层次上对语义相关性建模。

1.3. 本文组织结构

本文共分为五章，下面是余下各章的简介。

第二章介绍本文的相关工作及研究背景。我们首先回顾了图像语义自动标注中的大量工作，分别对生成模型与判别模型进行分析。其次，我们回顾了物体识别以及图像语义自动标注中对语义上下文建模的相关工作。最后，我们列举了 MRF 和 CRF 在计算机视觉上的相关应用。

第三章描述了多马尔科夫随机场上下文相关模型。我们具体介绍了 MMRF 模型图结构的构造，基于生成模型的势函数设计，正则化 (Regularization) 最大伪似然 (Pseudo-likelihood) 参数估计以及模型推理算法。此外，我们在 Corel 数据集以及 TRECVID-2005 数据集上进行实验，并给出实验结果。

第四章提出了最大边缘条件随机场上下文相关模型。我们具体描述了 MMCRF 模型图结构的构造，基于线性判别模型的势函数设计，最大边缘参数估计以及 MMCRF 的模型推理。我们给出了 MMCRF 模型在 Corel 数据集以及 TRECVID-2005 数据集上的实验结果，并与相关方法进行了比较。

第五章总结了本文的研究工作。我们回顾了本文的主要贡献，并展望了将来进一步的工作。

第二章 相关工作及研究背景

2.1. 图像语义自动标注

由于图像语义自动标注在基于关键词的图像视频检索上有着巨大的应用前景，它得到了研究者越来越多的关注。现有的图像语义自动标注方法可大体上分为生成模型与判别模型两大类。生成模型通过估计图像与关键词共同出现的联合概率分布来进行图像自动标注。一个开创性的工作是 Duygulu 等提出的机器翻译模型[8]。该模型首先把图像分割成区域，再把图像区域聚类得到不同的区域类型，最后采用机器翻译的方法学习从图像区域类型到关键词的映射。此外，[8]发布了一个包含 5000 幅图像和 374 个关键词的 Corel 数据集，现已成为图像语义自动标注研究中的基准数据集。采用生成模型处理图像语义自动标注的另一个突破性进展是相关模型的提出[1, 2, 3]。交叉媒体相关模型 (CMRM) [1]采用图像区域聚类后的区域类型来表示图像，再通过统计图像区域类型以及关键词在已标注图像中的出现频率来估计该图像从已标注图像生成的概率。连续空间相关模型 (CRM) [2]对图像区域采用非参数估计来估计图像区域的生成概率，对关键词采用多项式分布来估计关键词的生成概率。CRM 是对 CMRM 的改进。多伯努利相关模型 (MBRM) [3]采用多伯努利分布来估计关键词的生成概率，是对 CRM 的改进。在相关模型的基础上，研究者们通过在标注模型中利用关键词之间的相关性来提高图像标注的性能。Liu 等[9]提出了对偶交叉媒体相关模型 (Dual Cross-Media Relevance Model, DCMRM)。DCMRM 通过对预先定义的词典中的关键词求期望来估计图像与关键词共同出现的联合概率，从而利用关键词之间的相关性，图像检索技术以及 web 搜索技术来推断图像的语义。Zhou 等[10]提出了一个迭代式图像语义标注算法。该算法利用关键词之间的相关性和图像区域匹配来提高图像标注的性能。Jin 等[39]提出了连贯语言模型 (Coherence Language Model, CLM)。CLM 利用 EM 算法来寻找给定图像最优的语言模型，并在 EM 算法中利用了关键词之间的相关性。Wang 等[19]提出了一个基于马尔科夫链的图像标注方法。该方法把关键词看作是马尔科夫链的状态，从而能够利用关键词之间的关系来改进标注性能。除了从训练数据中获取语义概念之间的相互关系外，一些工作利用现有的本体 (Ontology) 来度量语义概念之间的相关性，例如 Google 搜索引擎[44]和 WordNet [45, 46]。相关模型外的另一大类生成模型是主题模型 (Topic Model)。这类方法把主题看作是图像视觉特征和语义概念上的联合概率分布，而已标注的图像是某一特定主题的样本。典型的主题模型有隐藏 Dirichlet 分配 (Latent Dirichlet Allocation, LDA) 模型[55, 59]，概率隐藏语义

分析 (Probabilistic Latent Semantic Analysis, PLSA) 模型[56]和层次 Dirichlet 过程 (Hierarchical Dirichlet Process, HDP) 模型[57]。

判别模型把每个语义概念看作一个类, 采用分类技术来解决图像语义自动标注问题。早期的工作关注于从图像中区分特定的语义概念, 例如区分室内场景和室外场景[4], 区分城市景色和山水风光[5], 检测图像中的树木[49], 马匹[50], 或建筑[51], 等等。这些工作把图像自动标注看作二分类问题。然而我们在图像自动标注中常常面对的是多个语义概念, 把图像自动标注看作多分类问题更为确切。Cusano 等[6]利用多分类支持向量机 (SVM) 把图像区域分类到预先定义的七个类。Yang 等[20]提出了基于非对称 SVM 的多样例学习算法 (ASVM-MIL), 在多样例学习 (Multiple-Instance Learning) [60]框架下, 采用 SVM 来处理图像语义自动标注。Carneiro 等[7]提出了有监督多分类标注 (SML) 模型。SML 采用高斯混合模型同时对单幅图像以及具有相同标注的同一类图像建模来估计类间概率。与在生成模型中利用语义概念之间的相关性类似, 在判别模型中利用不同类别之间的相互关系能够改进图像标注性能。Wang 等[47]采用局部多标签分类 (Multi-label Classification) 技术来进行图像语义自动标注, 从而在分类过程中利用语义概念之间的关系。Fan 等[52]采用层次分类 (Hierarchical Classification) 技术来实现多层次的图像语义自动标注, 同时在标注过程中利用了类别间的层次关系。

除标注模型的能力外, 图像的视觉特征对图像语义自动标注的性能起着至关重要的作用。Makadia 等[21]基于 K 近邻算法提出了图像语义自动标注的一个基准方法。该方法提取了图像的多种颜色和纹理特征, 并在每一种特征上计算距离, 而两幅图像间的距离是多种特征上距离的平均值 (称为 Joint Equal Contribution, JEC)。最后标注方法基于所计算出的距离采用标签传播 (label propagation) 算法进行图像标注。该方法在公共的语义标注数据集上表现出了很强的竞争力, 展示了图像视觉特征对于语义自动标注性能的影响。随后, Guillaumin 等[22]采用了更为丰富的多种视觉特征, 并利用基于最近邻的判别度量学习算法来自动学习各种特征距离上的权重, 取得了优于[21]的标注性能。Zhang 等[58]提出了一种被称作是群稀疏 (Group Sparsity) 的正则化方法用于特征选择。该方法把每一种特征看作一个群, 群内采用 L2 正则化, 而群间采用 L1 正则化, 从而同时利用了特征的稀疏性和特征群的内部特性。Wang 等[48]在交叉媒体相关模型中把图像的全局特征, 区域特征以及上下文特征结合起来, 改进了图像语义标注的性能。

2.2. 语义上下文建模

对人类视觉系统的研究表明, 场景中物体之间的相互关系会影响人们物体识别的

效率[23]。上下文建模 (Context Modeling) 在物体识别以及图像语义自动标注中都得到人们越来越多的重视。根据图像在上下文相关模型中的表示形式, 我们大致把它们分成基于物体的上下文相关模型 (object-based contextual model) 和基于场景的上下文相关模型 (scene-based contextual model)。基于物体的上下文相关模型首先把图像分割成区域, 每个区域代表一个物体, 再对物体之间的相互关系建模。例如, [24]和[25]通过考虑物体在场景中的相对位置, 对物体间的空间上下文建模。[26]和[53]利用物体在场景中共同出现的信息来对物体的语义上下文建模。基于场景的上下文相关模型把图像看作一个整体, 利用图像场景的统计信息作为上下文信息。在现有的工作中, 有两种主要的图像整体表示形式。一种是 Oliva 和 Torralba 等在[27]提出的, 通过图像的二阶统计量对场景建模, 提出了 Gist 特征。另一种被称为 “bag-of-features” (BOF) 表示形式。BOF 首先提取图像底层的局部特征, 再把这些局部特征聚合起来形成对图像场景的整体描述。例如, Lowe [42]提出的 SIFT 特征是 BOF 特征的一个典型代表。Li 和 Perona [28]采用 BOF 表示形式来学习和识别自然场景的类别。Lazebnik [54]等在 BOF 特征表示的基础上引入特征的空间位置信息, 增强了视觉特征的判别能力。在图像语义自动标注中, 由于图像的整体表述形式具有应用到大规模的图像视频检索上的潜力, 基于场景的上下文相关模型受到了更多的关注。语义上下文, 即语义概念之间的共同出现, 是图像语义自动标注中的主要上下文信息。

在图像语义自动标注中, 现有的语义上下文相关模型从两个不同的角度来利用语义上下文信息。第一类模型利用语义上下文对现有标注方法的结果进行改进。Rasiwasia 和 Vasconcelos [11]在 SML 标注模型[7]的基础上, 采用混合 Dirichlet 分布构造语义层模型。从而利用语义概念之间的相关性来改进 SML 的标注结果。我们在本文中提出的多马尔科夫随机场上下文相关模型在 MBRM [3]的基础上, 利用 MRF 构造语义层模型来改进 MBRM 的标注结果。第二类模型把语义上下文建模看作是结构化分类 (Structural Classification) 问题。CML 模型[12]基于结构化 SVM [13], 把一幅图像同时标注上多个语义概念, 并在标注过程中利用语义概念之间的相关性。我们提出的最大边缘条件随机场上下文相关模型把结构化分类问题巧妙地转化为多个相互联系的二分类问题, 从而避免求解结构化 SVM 中约束数目为语义概念数目指数级的最优化问题。

2.3. 马尔科夫随机场

马尔科夫随机场 (MRF) 理论是概率论的一个分支, 用于对随机变量及它们之间的相互关系建模。MRF 广泛应用于计算视觉的多个领域, 从底层的视觉问题, 例如图像恢复[29]和纹理建模[30], 到高层的视觉问题, 例如图像分割[31]和物体检测[32]。

在这些应用中，MRF 用于对图像像素或图像区域间的空间关系建模。而我们在本文中提出的图像语义自动标注框架是利用 MRF 对语义概念之间的相互关系建模，这是本文工作的主要创新。

MRF 理论在计算机视觉领域的进一步应用得益于条件随机场 (CRF) [18] 的提出。MRF 通常与贝叶斯法则一起用于估计观察数据与该数据标签的联合概率分布。在这个框架下，人们通常假设在已知标签的情况下，观察数据是相互独立的。然而不少研究者指出这个假设在大多数情况下过于严格。因此，Lafferty 等 [18] 提出了条件随机场模型，从判别分析的角度直接对标签的后验概率建模。CRF 起初用于分割和标记一维的序列数据，并且 CRF 能够捕获观察数据间任意的依赖关系。随后 CRF 被广泛应用于多种计算视觉问题，例如图像分割 [33]，物体识别 [34] 和图像集合标注 [35]。CRF 的二维版本被称作是判别随机场 (Discriminative Random Field, DRF) [36]。Kumar 和 Hebert 通过 DRF 对图像区域之间的空间关系建模。MRF 领域的另一个重要工作是 Taskar 等提出的最大边缘马尔科夫网络 (Max-Margin Markov Network, M3N) [37]。M3N 在最大边缘框架下，通过 MRF 来利用变量之间的相互关系，把需要求解的最优化问题的约束数目从指数级降低到多项式级。我们在本文中提出的最大边缘条件随机场模型与 M3N 的不同之处在于，MMCRF 引入可拆分的 Hinge 损失函数，通过求解一系列独立的 QP 问题来训练具有最大边缘的 CRF 模型。

第三章 多马尔科夫随机场上下文相关模型

我们在本章中提出了多马尔科夫随机场（MMRF）上下文相关模型。MMRF 在底层生成模型的基础上，为每一个语义概念构造语义层的 MRF 模型，从而利用语义概念之间的相关性来提高图像语义自动标注的性能。与在生成模型中引入语义概念相关性的工作[9, 10]不同，我们通过 MRF 引入了适当的参数对语义上下文建模，并针对不同的语义概念学习相应的 MRF，大大增强了模型的学习能力，使模型能够准确地捕获语义概念之间的相互关系。具体来讲，我们首先根据语义概念在训练数据集中共同出现的频率，为每一个语义概念构造 MRF 的图结构。其次，我们利用生成模型估计图像与语义概念共同出现的联合概率，并把这些联合概率当作是 MRF 每个点上的观察值。从而，我们基于生成模型设计了 MMRF 的点势函数和边势函数。再次，我们通过正则化最大伪似然估计来学习每个 MRF 的参数。最后，我们利用迭代条件模式（Iterated Conditional Modes, ICM）[38]算法来进行模型推理。图 3.1 描绘了 MMRF 上下文相关模型的基本框架。从图中可明显看出，MMRF 模型是一个两层模型。上层是在语义空间中构造的 MRF 模型，下层是利用生成模型构造的概率空间。 $P(\mathbf{x}, w_i)$ 表示观察值 \mathbf{x} 与关键词 w_i 共同出现的联合概率，可通过图像语义自动标注中的某个生产模

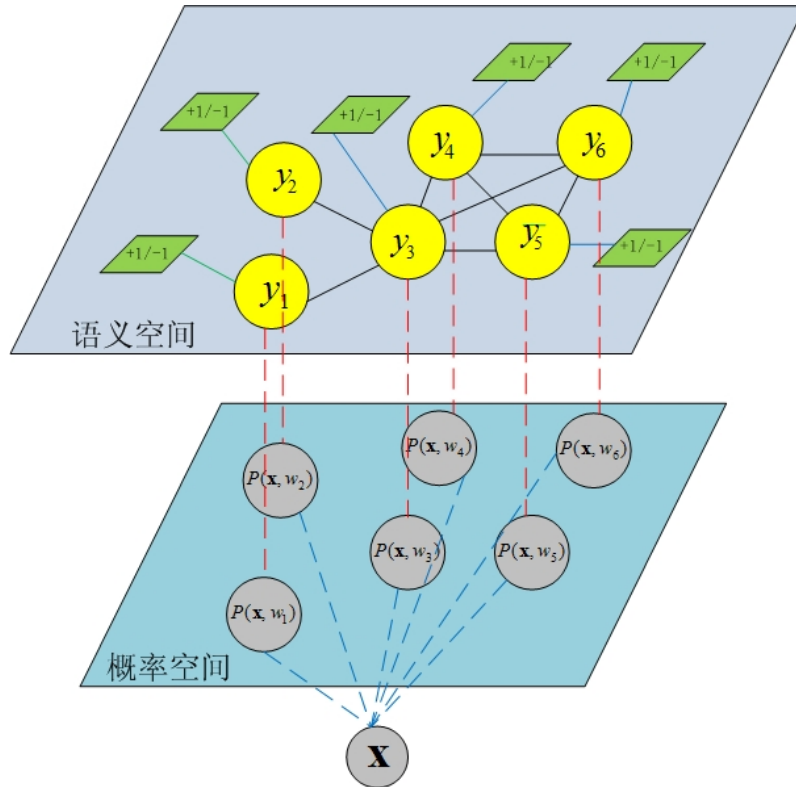


图 3.1 MMRF 上下文相关模型的基本框架

型估计得到。MMRF 把 $P(\mathbf{x}, w_i)$ 看作是 MRF 点 i 上的观察值。因此，MMRF 通过生成模型作为中介，把观察值与 MRF 的标签联系起来。

本章余下内容组织如下：3.1 节描述了 MMRF 图结构的构造；3.2 节给出了 MMRF 势函数的设计；3.3 节和 3.4 节分别介绍了 MMRF 的参数估计和模型推理；3.5 节总结了 MMRF 图像语义自动标注算法；3.6 节给出了 MMRF 图像语义自动标注的实验结果；最后在 3.7 节对本章进行小结。

3.1. 概念图

在我们的模型中，MRF 的图结构是基于语义概念在训练数据集中的共同出现来构造的。给定训练集 $T = \{(\mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^T$ ，其中 \mathbf{x}^t 表示第 t 幅图像的视觉特征， $\mathbf{y}^t = (y_1^t, y_2^t, \dots, y_{|S|}^t)$ 是第 t 幅图像的标签向量， $y_i^t \in \{-1, +1\}$, $i = 1, 2, \dots, |S|$ 表示第 i 个语义概念在图像中不出现或出现， $|S|$ 是语义概念的数目即 MRF 点集的大小， $|T|$ 表示训练集的大小。在训练集 T 中，每一幅图像对应多个语义概念。这一现象和文本检索中文本的“Bag-Of-Words”(BOW)表示形式相似。如果我们把每幅图像看作是一个文档，图像相关联的语义概念看作文档中的词，那么整个训练集就可以当作一个语料库。如果两个语义概念在该语料库中共同出现，我们则定义这两个语义概念相关。基于这样定义的语义概念之间的相关性，我们在语义概念上构造图 $G = (S, E)$ ，其中 $S = \{1, 2, \dots, m\}$ 对应 MRF 的点集， $(i, j) \in E$ 当且仅当语义概念 i 和 j 相关。

MMRF 模型为每一个语义概念构造自身的 MRF 来区分语义概念间的不同。我们从概念图 G 中抽取不同的子图来构造这些 MRF 的图结构。语义概念（关键词） w_i 对应的子图为 $G_i = (S_i, E_i)$ ，其中 $S_i = \{i\} \cup N_i$ ， $E_i = \{(i, j) | i, j \in S_i \text{ 且 } (i, j) \in E\}$ ，即 G_i 是由点 i 以及它的邻居构成的子图。因此，我们采用与关键词 w_i 相关的所有关键词来构造 w_i 的 MRF 模型。在本章后面的内容中，我们具体描述针对某个关键词的 MRF。为了使描述更为简洁，我们仍然采用 S 来表示 MRF 的点集。

3.2. 基于生成模型的势函数设计

利用图像语义自动标注中的生成模型，例如 CRM[2]和 MBRM[3]，我们可以为每一个关键词估计它与观察图像共同出现的联合概率。在 MMRF 中，我们把关键词 w_i 与观察图像 \mathbf{x} 共同出现的联合概率 $P(\mathbf{x}, w_i)$ 看作是 MRF 点 i 上的观察值。我们设计点势函数的动机是当 $P(\mathbf{x}, w_i)$ 较大时，相应的标签 y_i 应该为 +1；反之， y_i 取 -1。我们把点势函数定义为

$$V_i(y_i) = y_i (\lambda_i + \alpha_i P(\mathbf{x}, w_i)), \quad (3.1)$$

其中 λ_i 和 α_i 是需要估计的参数。从公式(3.1)中可知，当 $\alpha_i < 0$ 时，高联合概率 $P(\mathbf{x}, w_i)$ 与正标签结合比与负标签结合产生的势能小。因此， $P(\mathbf{x}, w_i)$ 越大， y_i 取 +1 的概率也就越大。MMRF 的边势函数定义为

$$V_2(y_i, y_j) = \beta_{ij} y_i y_j P(\mathbf{x}, w_j), \quad (3.2)$$

其中 β_{ij} 是待估计的参数。我们通过边势函数引入与点 i 相邻的点 j 上的标签和联合概率对点 i 的影响。将等式(3.1)和(3.2)代入等式(1.3)中，我们得到 MMRF 的能量函数：

$$U(\mathbf{y} | \boldsymbol{\theta}) = \sum_{i \in S} y_i (\lambda_i + \alpha_i P(\mathbf{x}, w_i)) + \sum_{i \in S} \sum_{j \in N_i} \beta_{ij} y_i y_j P(\mathbf{x}, w_j), \quad (3.3)$$

其中 $\boldsymbol{\theta} = \{\alpha_i, \beta_{ij} | i \in S, j \in N_i\}$ 表示 MMRF 的参数。根据 Hammersley-Clifford 定理[17]，标签 \mathbf{y} 的联合概率分布可表示为

$$P(\mathbf{y}) = Z^{-1} \times e^{-U(\mathbf{y} | \boldsymbol{\theta})}, \quad (3.4)$$

其中

$$Z(\boldsymbol{\theta}) = \sum_{\mathbf{y} \in Y} e^{-U(\mathbf{y} | \boldsymbol{\theta})} \quad (3.5)$$

是划分函数。为了利用等式(3.4)来求解 MRF 概率最大的标签组合，我们首先需要估计出参数 $\boldsymbol{\theta}$ ，我们在下一节中介绍 MMRF 的参数估计。

3.3. 正则化最大伪似然参数估计

MRF 中最常用的参数估计方法是最大似然估计。最大似然估计选择使训练数据标签的联合概率(3.4)最大的参数。在计算标签的联合概率时需要计算划分函数(3.5)的值。这在实际问题中通常是不可行的，因为标签组合的数目是标签数目的指数倍。在 MMRF 模型中，我们采用被称为最大伪似然参数估计的近似参数估计方法来避免求解划分函数。

伪似然 (pseudo-likelihood) 定义为

$$PL(\mathbf{y}) = \prod_{i \in S} P(y_i | y_{N_i}) = \prod_{i \in S} \frac{e^{-U_i(y_i, y_{N_i})}}{\sum_{y_i} e^{-U_i(y_i, y_{N_i})}}, \quad (3.6)$$

其中

$$U_i(y_i, y_{N_i}) = V_1(y_i) + \sum_{j \in N_i} V_2(y_i, y_j) \quad (3.7)$$

可以看作是点 i 引入的势能。因为 y_i 和 y_{N_i} 并不是独立的，所以伪似然(3.6)不等于真似然。将等式(3.1)和(3.2)代入等式(3.7)中可得

$$U_i(y_i, y_{N_i}) = y_i (\lambda_i + \alpha_i P(\mathbf{x}, w_i)) + \sum_{j \in N_i} \beta_{ij} y_i y_j P(\mathbf{x}, w_j). \quad (3.8)$$

我们令

$$\boldsymbol{\theta}_i = (\lambda_i, \alpha_i, \beta_{ij \forall j \in N_i})^T \quad (3.9)$$

$$\mathbf{x}_i = (1, P(\mathbf{x}, w_i), y_j P(\mathbf{x}, w_j)_{\forall j \in N_i})^T, \quad (3.10)$$

则等式(3.7)可改写为

$$U_i(y_i, y_{N_i}) = y_i \boldsymbol{\theta}_i^T \mathbf{x}_i, \quad (3.11)$$

其中 $\boldsymbol{\theta}_i$ 可看作是点 i 上的参数, \mathbf{x}_i 是为点 i 构造的训练数据。将等式(3.11)代入等式(3.6)中可将伪似然表示为

$$PL(\mathbf{y}) = \prod_{i \in S} \frac{e^{-y_i \boldsymbol{\theta}_i^T \mathbf{x}_i}}{e^{-\boldsymbol{\theta}_i^T \mathbf{x}_i} + e^{\boldsymbol{\theta}_i^T \mathbf{x}_i}}. \quad (3.12)$$

我们需要估计的参数为 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \dots, \boldsymbol{\theta}_{|S|}^T)^T$ 。

我们针对当前需要训练的 MRF 构造训练集 $T = \{(\mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^{|T|}$, 其中 $\mathbf{x}^t = \{\mathbf{x}_1^t, \mathbf{x}_2^t, \dots, \mathbf{x}_{|S|}^t\}$ 是第 t 幅图像的视觉特征, 它由每个点上的训练数据(3.10)组成, $\mathbf{y}^t = (y_1^t, y_2^t, \dots, y_{|S|}^t)^T$ 是第 t 幅图像的标签, 它由每个点上的标签组成。MRF 在训练集 T 上的伪似然为

$$\prod_{t=1}^{|T|} PL(\mathbf{y}^t) = \prod_{t=1}^{|T|} \prod_{i \in S} P(y_i^t | y_{N_i}^t) = \prod_{i \in S} \prod_{t=1}^{|T|} P(y_i^t | y_{N_i}^t) = \prod_{i \in S} PL_i, \quad (3.13)$$

其中我们定义

$$PL_i = \prod_{t=1}^{|T|} P(y_i^t | y_{N_i}^t) \quad (3.14)$$

可看作是点 i 上的伪似然。需要注意的是, 每两个 PL_i 之间并没有相同的参数。因此, 最大伪似然参数估计 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \dots, \boldsymbol{\theta}_{|S|}^T)^T$ 可以通过分别最大化 PL_i 来求解 $\boldsymbol{\theta}_i$ 得到。这个性质不仅大大提高了参数估计的速度, 还使得我们可以针对不同的点来构造不同的训练集估计相应的参数。为每一个点构造自己的训练数据可以减轻训练集中数据不平衡的问题。下面我们通过最大化 PL_i 来求解 $\boldsymbol{\theta}_i$ 。

假设我们为点 i 构造了训练集 $T_i = \{(\mathbf{x}_i^t, y_i^t)\}_{t=1}^{|T_i|}$, 那么点 i 上的对数伪似然为

$$\begin{aligned} \ln PL_i &= \sum_{t=1}^{|T_i|} \ln P(y_i^t | y_{N_i}^t) \\ &= \sum_{t=1}^{|T_i|} \left\{ (1 - y_i^t) \boldsymbol{\theta}_i^T \mathbf{x}_i^t - \ln(1 + e^{2\boldsymbol{\theta}_i^T \mathbf{x}_i^t}) \right\}. \end{aligned} \quad (3.15)$$

当训练数据不足时, 大量的参数会导致过拟合 (overfitting)。为了解决这一问题, 我们采用高斯先验对等式(3.15)中的对数伪似然进行惩罚, 可得到下面的正则化对数伪似然:

$$L_i(\boldsymbol{\theta}_i) = \sum_{t=1}^{|T_i|} \left\{ (1 - y_i^t) \boldsymbol{\theta}_i^T \mathbf{x}_i^t - \ln(1 + e^{2\boldsymbol{\theta}_i^T \mathbf{x}_i^t}) \right\} - \frac{\|\boldsymbol{\theta}_i\|^2}{2\sigma^2}, \quad (3.16)$$

其中常数 σ 的值根据经验来选取，并且我们约束每个点上的 σ 都相同。为了最大化 (3.16)，我们令 $L_i(\boldsymbol{\theta}_i)$ 关于 $\boldsymbol{\theta}_i$ 的导数为零，可得到下面的评分方程：

$$\frac{\partial L_i(\boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i} = \sum_{t=1}^{|T_i|} \left\{ \mathbf{x}_i^t (1 - y_i^t - 2P(\mathbf{x}_i^t; \boldsymbol{\theta}_i)) \right\} - \frac{\boldsymbol{\theta}_i}{\sigma^2}, \quad (3.17)$$

其中

$$P(\mathbf{x}_i^t; \boldsymbol{\theta}_i) = \frac{e^{2\boldsymbol{\theta}_i^T \mathbf{x}_i^t}}{1 + e^{2\boldsymbol{\theta}_i^T \mathbf{x}_i^t}}. \quad (3.18)$$

我们采用 Newton-Raphson 算法来求解方程 (3.17)。 $L_i(\boldsymbol{\theta}_i)$ 的 Hessian 矩阵为

$$\frac{\partial^2 L_i(\boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i^T} = -4 \sum_{t=1}^{|T_i|} \left\{ \mathbf{x}_i^t \mathbf{x}_i^{tT} P(\mathbf{x}_i^t; \boldsymbol{\theta}_i) (1 - P(\mathbf{x}_i^t; \boldsymbol{\theta}_i)) \right\} - \frac{\mathbf{I}}{\sigma^2}, \quad (3.19)$$

其中 \mathbf{I} 是单位矩阵。假设当前参数值为 $\boldsymbol{\theta}_i^{\text{old}}$ ，Newton-Raphson 算法的迭代更新规则为

$$\boldsymbol{\theta}_i^{\text{new}} = \boldsymbol{\theta}_i^{\text{old}} - \left(\frac{\partial^2 L_i(\boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i^T} \right)^{-1} \frac{\partial L_i(\boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i}, \quad (3.20)$$

其中导数值是根据 $\boldsymbol{\theta}_i^{\text{old}}$ 来计算的。由于正则化对数伪似然是凹函数，Newton-Raphson 算法最终将达到收敛。通过分别对每个点进行正则化最大伪似然参数估计，我们可得到 MRF 的参数。

3.4. 模型推理

MRF 中模型推理的目的是为了寻找联合概率最大的标签组合：

$$\mathbf{y}^* = \arg \max_{\mathbf{y}} P(\mathbf{y}), \quad (3.21)$$

其中 $P(\mathbf{y})$ 在等式 (3.4) 中定义。我们利用迭代条件模式 (Iterated Conditional Modes, ICM) [38] 算法来进行模型推理。ICM 算法在迭代过程中通过最大化局部条件概率来顺序地更新标签。在 $(t+1)$ 步迭代中，给定图像视觉特征 \mathbf{x} 和除点 i 外其他点上的标签，ICM 通过最大化条件概率 $P(y_i | \mathbf{x}, y_{S-i}^{(t)})$ 顺序地将 $y_i^{(t)}$ 更新为 $y_i^{(t+1)}$ 。在 MRF 中， y_i 仅仅与它邻域内的标签相关。因此，最大化 $P(y_i | \mathbf{x}, y_{S-i}^{(t)})$ 等价于最大化 $P(y_i | \mathbf{x}, y_{N_i}^{(t)})$ 。我们可通过最小化相应的势函数得到下面的更新规则：

$$y_i^{(t+1)} \leftarrow \arg \min_{y_i} U_i(y_i, y_{N_i}^{(t)}), \quad (3.22)$$

其中 $U_i(y_i, y_{N_i}^{(t)})$ 在等式 (3.11) 中定义。上面的更新规则等价于

$$y_i^{(t+1)} = \begin{cases} +1, & \text{if } \theta_i^T \mathbf{x}_i \leq 0 \\ -1, & \text{if } \theta_i^T \mathbf{x}_i > 0 \end{cases}, \quad (3.23)$$

其中 θ_i 是点 i 上估计好的参数， \mathbf{x}_i 是基于图像的视觉特征为点 i 构造的训练数据。从一个初始标签组合开始（所有标签设置为 -1 ），ICM 迭代到算法收敛。我们则得到了近似的联合概率最大的标签组合。

3.5. MMRF 图像语义自动标注算法

3.5.1. 训练集的构造

在 3.3 节中，最大正则化伪似然参数估计需要为 MRF 的每一个点构造自己的训练集。假设我们需要为点 i 构造训练集 T_i ，而点 i 对应的关键词为 w_i 。我们把训练集中标注为 w_i 的图像称为 w_i 的正样例，未标注 w_i 的图像称为 w_i 的负样例。首先，我们从整体训练集 T 中采样得到一个 w_i 的正负样例数目更为平衡的训练集 T'_i 。在现实的真实数据中，负样例通常大大多于正样例，采样技术能够减轻数据不平衡带来的影响。我们选取关键词的所有正样例，并随机选取负样例的一个子集。所选取的负样例数目比正样例数目多一个小的因子 δ 。在实验中，我们令 $\delta=1$ 。当训练集中的正样例足够多时，这多出的负样例对于模型训练的影响不太。而由于正样例数目不足，不能准确捕获关键词的语义时，多出的负样例能够避免模型在预测时产生过多错误的正样例。其次，

算法 3.1 训练集构造算法

1. **输入：** 整体训练集 T ，工作 MRF MRF
2. **输出：** MRF 的训练集 T''
3. **for** MRF 的每一个点 i **do**
4. 对 T 采样得到正负样例更为平衡的训练集 T'_i
5. **for** 每一个 $\mathbf{x}^t \in T'_i$ **do**
6. 提取出标签 y_i^t 和 $y_j^t, j \in N_i$
7. 计算联合概率 $P(\mathbf{x}^t, w_i)$ 和 $P(\mathbf{x}^t, w_j), j \in N_i$
8. 计算 $\mathbf{x}_i^t = (1, P(\mathbf{x}^t, w_i), y_j^t P(\mathbf{x}^t, w_j)_{j \in N_i})^T$
9. **end for**
10. $T_i = \{(\mathbf{x}_i^t, y_i^t)\}_{t=1}^{|T'_i|}$
11. **end for**
12. $T'' = \bigcup_{i=1}^{|S|} T_i$

算法 3.2 多马尔科夫随机场（MMRF）图像语义自动标注算法

1. **输入：**待标注图像 I ，关键词词表 V ，训练数据集 T ，关键词图 G
2. **输出：**图像 I 的语义标注
3. **for** 每一个 $w \in V$ **do**
4. 从图 G 中抽取出子图 G_w 作为 MRF_w 的图结构
5. 根据算法 3.1 为 MRF_w 构造训练集 T_w''
6. 基于 T_w'' 估计 MRF_w 的参数
7. 在 MRF_w 上推理得到 I 关于 w 的标签
8. **end for**

对于 T_i' 中的每一幅图像 \mathbf{x}' ，我们提取出点 i 及其邻域 N_i 上的标签 y_i' 和 y_j' , $j \in N_i$ ，并利用生成模型计算这些点上的联合概率 $P(\mathbf{x}', w_i)$ 和 $P(\mathbf{x}', w_j)$, $j \in N_i$ 。最后，我们根据等式 (3.10) 将这些标签与联合概率组合起来得到 $T_i = \{(\mathbf{x}_i', y_i')\}_{i=1}^{|T_i'|}$ 。算法 3.1 描述了 MMRF 模型训练集构造的算法。

3.5.2. 标注算法

在构造好的训练集上为每一个关键词训练好 MRF 后，我们通过 MRF 模型推理来对图像进行语义自动标注。对于一幅输入图像 I ，每一个 MRF 都会输出一个标签向量。但我们仅仅把 w_i 对应的 MRF 中 w_i 的标签作为图像 I 的标注词。因此，我们需要在每一个 MRF 上推理后才能得到图像的最终标注。算法 3.2 描述了采用多马尔科夫随机场（MMRF）模型进行图像语义自动标注的过程。需要注意的是，如果我们的任务是标注一组图像，那么算法中每一个 MRF 关键词子图的构造，训练集的构造以及参数估计都只需要执行一次。

3.6. 实验

3.6.1. 实验数据集

Corel 数据集：我们采用 Corel 数据集[8]来进行实验。该数据集被广泛应用于图像语义自动标注的性能比较。Corel 数据集包含 5000 幅图像，其中 4500 幅是训练图像，其余 500 幅图像用于测试。每幅图像被标注了 1-5 个语义关键词，而数据集中一共有 374 个关键词。数据集中每幅图像被分割为 1-10 块区域，每块区域提取了 36 维的特征向量，包括区域的颜色，纹理以及区域位置信息[8]。除了区域特征外，CRM[2]和

MBRM[3]采用网格特征取得了更好的标注性能。因此，我们对 Corel 数据集中的图像提取了一种新的网格特征。每幅图像被划分成 26 个矩形网格，包括 5×5 的网格和一个图像中央的网格。我们对每个网格提取了 528 维的特征向量，包括 448 维的颜色特征（局部和全局的颜色直方图）和 80 维根据 MPEG7 提取的纹理特征。在实验中，我们采用区域特征和网格特征来测试模型的性能。我们在模型的名称后面加上“-grid”表示我们采用了网格特征。例如，MBRM-grid 表示采用了我们的网格特征的 MBRM 方法。

TRECVID-2005 数据集：为了测试 MMRF 对于视频关键帧标注的性能，我们在 TRECVID-2005 标准数据集上进行实验。该数据集包含有 108 小时的多种语言的新闻视频。这些视频被切割为 61901 帧。每一帧图像被进一步分割为 5×5 的网格，每一个网格提取了 9 维的视觉特征向量。这相当于每一帧图像提取了一个 225 维的特征向量。在 TRECVID-2005 数据集中包含有 39 个关键词，而每一帧被标注了 0-11 个关键词。在所有的数据上进行实验是很花费时间的。因此，我们随机选取了 9000 帧图像作为训练数据，随机选取了另外 1000 帧图像作为测试数据。所选取的每一帧图像至少被标注了一个关键词。

3.6.2. 评价度量

和先前图像语义自动标注的工作相同，我们采用查全率(recall)和查准率(precision)来度量标注性能。给定一个查询关键词 w ，令 $|W_G|$ 表示测试集中人工标注为 w 的图像的数目， $|W_M|$ 表示标注算法标注为 w 的测试图像的数目， $|W_C|$ 表示标注算法标注正确的测试图像的数目，则查全率与查准率分别定义为：

$$recall = \frac{|W_C|}{|W_M|}, \quad precision = \frac{|W_C|}{|W_G|}.$$

我们分别对每个关键词计算查全率和查准率，再对每个关键词的查全率和查准率求平均作为标注性能的评价度量。

3.6.3. 在 Corel 数据集上的对比

因为 MBRM[3]是图像语义自动标注中一个具有代表性的生成模型，且 MBRM 取得了很有竞争力的标注性能，我们首先在 Corel 数据集上将 MMRF 与 MBRM 进行对比。MMRF 采用 MBRM 估计的图像视觉特征与语义关键词共同出现的联合概率来构造 MRF。由于 MBRM 不能自动地决定每幅图像应该标注的关键词的数目，我们把 MBRM 对每幅图像的标注长度设置为 5 个关键词。我们提出的 MMRF 模型能够自动

表 3.1 在 Corel 数据集上采用区域特征与 MBRM 的对比

模型	MBRM	MMRF
查全率大于 0 的关键词数目	109	124
图像的平均标注长度	5	4.3
263 个关键词上的结果		
平均查全率	0.20	.023
平均查准率	0.19	0.27
性能最好的 49 个关键词上的结果		
平均查全率	0.68	0.67
平均查准率	0.64	0.76

决定每幅图像的标注长度。表 3.1 列出了 MMRF 与 MBRM 在 Corel 数据集上采用区域特征的标注性能对比。从表格中可以看出,我们的 MMRF 模型与 MBRM 模型相比,明显地改进了 Corel 数据集上的图像语义自动标注性能。对于测试集中出现的 263 个关键词,MMRF 在平均查全率上取得了 15%的改进,同时在平均查准率上取得了 42%的改进。对于标注性能最好的 49 个关键词(F1 最大的 49 个关键词),MMRF 在平均查准率上取得了 19%的改进,同时取得了与 MBRM 接近的平均查全率。我们的方法对于每幅图像平均标注了 4.3 个关键词,小于 MBRM 的平均每幅图像 5 个关键词。此外,MMRF 具有 124 个查全率大于 0 的关键词,多于 MBRM 的 109 个关键词。这说明了我们的方法对于正样例数目较少的关键词能够比 MBRM 取得更好的标注性能。

采用网格特征 MMRF 和 MBRM 都能够取得比区域特征更好的标注性能。表 3.2 列出了 MMRF 与 MBRM 在 Corel 数据集上采用网格特征的标注性能对比。对于测试

表 3.2 在 Corel 数据集上采用网格特征与 MBRM 的对比

模型	MBRM	MMRF	SMRF
查全率大于 0 的关键词数目	123	172	136
图像的平均标注长度	5	5.2	9.6
263 个关键词上的结果			
平均查全率	0.25	0.36	0.28
平均查准率	0.23	0.31	0.20
性能最好的 49 个关键词上的结果			
平均查全率	0.75	0.79	0.69
平均查准率	0.73	0.80	0.63

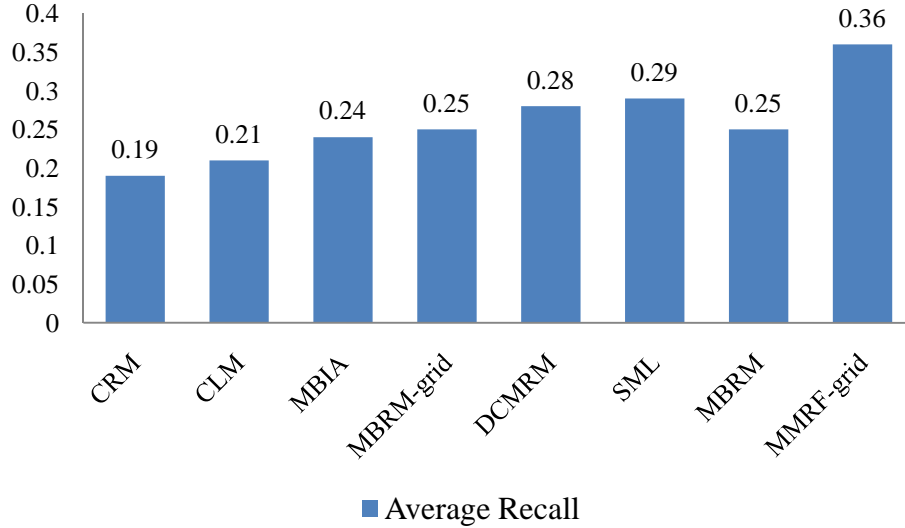


图 3.2 MMRF 与其他标注方法在平均查全率上的对比

集中的 263 个关键词，MMRF 有 172 个关键词的查全率大于 0，相比于 MBRM 取得了 40% 的显著改进。MMRF 在这 263 个词上的平均查全率和平均查准率分别为 0.36 和 0.31，相比于 MBRM 分别取得了 44% 和 35% 的显著改进。对于标注性能最好的 49 个关键词，MMRF 的平均查全率与平均查准率都超过了 MBRM。总体上讲，实验结果表明我们的方法能够大大改进语义自动标注的准确率，并且具有很强的能力来处理正样例数目少的关键词。我们在表 3.2 中同时给出了对于 Corel 数据集中的 374 个关键词只构造一个 MRF 进行标注的实验结果：SMRF (Single MRF)。SMRF 的标注性能远远低于 MMRF。这说明通过对每一个关键词构造一个 MRF，MMRF 模型能够准确地捕获到不同关键词的语义，同时避免了求解全局最优的参数设置，从而取得了更好的标注性能。

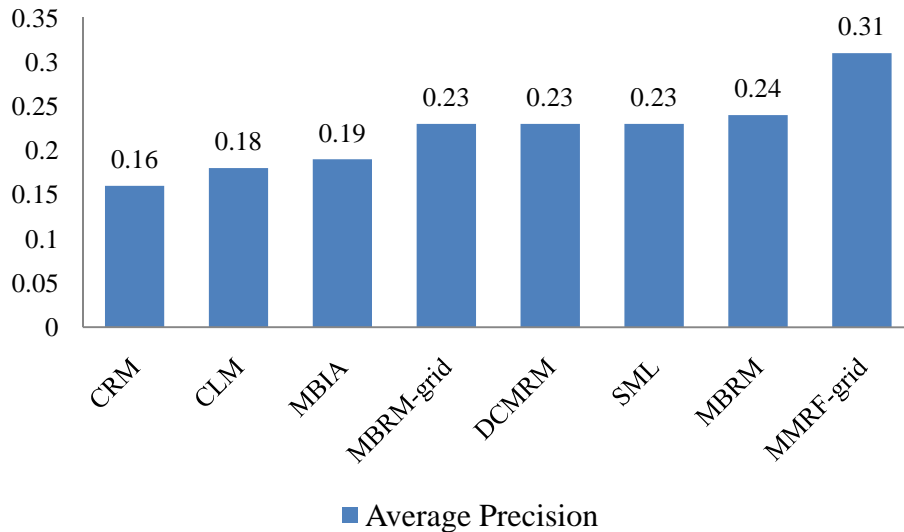







图 3.3 MMRF 与其他标注方法在平均查准率上的对比

表 3.3 Corel 数据集上一些标注的例子

					
MMRF-grid	leaf, flowers, petals, stems	grass, cars, tracks, prototype	grass, cow, bulls, antlers, elk, caribou	people, flowers, restaurant, shops, street, festival	light, shops
MBRM-grid	sky, water, flowers, bush, petals	water, grass, cars, tracks, prototype	sky, water, grass, antlers, elk	water, flowers, display, shops, street	sky, water, tree, light, shops
人工标注	leaf, flowers, petals, stems	cars, tracks, turn, prototype	cow, bulls, antlers, elk	tree, people, restaurant, tables	light, shops

除了 MBRM 外，我们还把 MMRF 与其它五种最新的图像语义自动标注方法在 Corel 数据集上进行比较。它们包括生成模型：CRM [2]，CLM [39]和 DCMRM [9]，判别模型：MBIA [19]和 SML [7]。图 3.2 和图 3.3 分别给出了 MMRF 与这些方法在 263 个关键词上平均查全率与平均查准率的比较。从图中可以看出，我们的方法取得了最高的平均查全率和平均查准率，并且与其他方法中性能最好的方法相比取得了超过 24% 的改进。

表 3.3 给出了 MMRF 与 MBRM 在 Corel 数据集上标注结果的一些例子。从表中可以看出，我们的方法不仅标注出了 MBRM 标注正确的关键词，而且比 MBRM 标注出了更多正确的关键词，同时减少了错误标注的数目。例如，MMRF 对于第一幅与第五幅图像的标注结果与人工标注结果完全相同，而 MBRM 有一些错误的标注词。对于第三幅图像，MMRF 甚至标注出了被人工标注忽略的正确的标注词“caribou”。

3.6.4. 在 TRECVID-2005 数据集上的对比

对于视频数据，我们把 MMRF 与 MBRM 在 TRECVID-2005 数据集上进行对比。MMRF 采用 MBRM 估计的图像视觉特征与语义关键词共同出现的联合概率来构造 MRF。我们把 MBRM 对于每帧视频图像的标注长度设置为 5，这个标注长度在我们的实验中取得了最好的性能。表 3.4 给出了实验结果。从表中可以看出，MMRF 预测出了 TRECVID-2005 数据集中的全部 39 个关键词，而 MBRM 仅仅预测出了 32 个。与 MBRM 相比，MMRF 在 39 个关键词上的平均查全率和平均查准率分别取得了 21% 和 42% 的改进。图 3.4 给出了 MMRF 与 MBRM 在 39 个关键词上性能的具体对比。从图

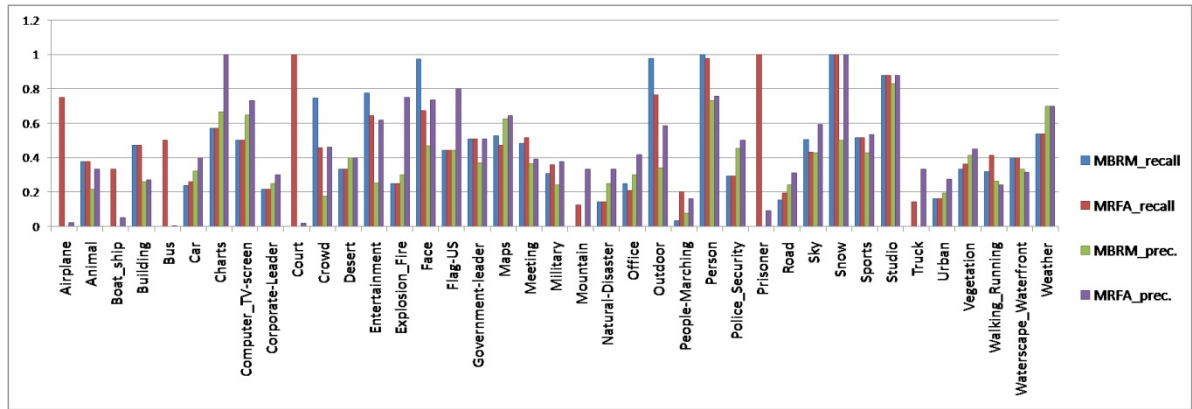


图 3.4 MMRF 与 MBRM 在 TRECVID-2005 数据集 39 个关键词上的性能对比

中可以看出，与 MBRM 相比，我们的方法在大多数关键词上对于查准率都取得了很大的改进。对于查全率来说，我们有 14 个关键词比 MBRM 高，有 17 个关键词与 MBRM 相同。MMRF 能预测出 MBRM 预测不出的正样例数目少的词，例如“Mountain”，“Prisoner”和“Truck”。

表 3.4 在 TRECVID-2005 数据集上与 MBRM 的对比

模型	MBRM	MMRF
查全率大于 0 的关键词数目	32	39
图像的平均标注长度	5	3.62
39 个关键词上的结果		
平均查全率	0.39	0.47
平均查准率	0.32	0.45

3.7. 本章小结

在本章中，我们提出了多马尔科夫随机场模型来增强图像语义自动标注中生成模型的学习能力。我们的方法能够适当地对语义上下文建模，从而利用语义概念之间的相关性来改进图像语义自动标注的性能。我们基于生成模型构造了 MRF 新颖的势函数，并高效地进行了模型的参数估计和推理。我们在公用的标准数据集上进行了实验。实验结果表明我们的方法具有很强的处理正样例少的关键词的能力，而且能够自动地决定每幅图像的标注长度。通过与当前最新的语义标注方法的比较，我们的方法对于标注性能做出了显著的改进。

第四章 最大边缘条件随机场上下文相关模型

我们在第三章中介绍了多马尔科夫随机场（MMRF）图像语义自动标注上下文相关模型。MMRF 在底层生成模型的基础上，构造高层的 MRF 来对语义上下文建模。两层的建模框架降低了模型的效率，并且标注性能的改进受到底层生成模型的限制。本章中，我们在第一章中介绍的 MRF 标注框架下，提出了最大边缘条件随机场（Maximal Margin Conditional Random Field, MMCRF）上下文相关模型来解决 MMRF 模型存在的问题。

我们提出的 MMCRF 模型是一个新颖的判别条件随机场（Conditional Random Field, CRF）[18]模型。MMCRF 直接根据图像的视觉特征来预测与图像相关的语义概念，并且在预测过程中能够利用语义概念之间的相关性。因此我们可以把 MMCRF 看作一个多标签分类模型。与基于物体的上下文相关模型中的 CRF [25, 26]不同，我们提出的 CRF 是在语义概念上构造的对语义概念相关性建模的模型。与先前的语义上下文相关模型[11, 12, 15]相比，MMCRF 模型能够同时从语义层次与视觉层次上对语义相关性建模。这个性质增强了模型对语义上下文建模的能力。具体来讲，基于第一章中提出的 MRF 标注框架，MMCRF 中的点表示语义概念，边表示两个语义概念相关。每个点上有一个二值标签表示相应的语义概念在图像中出现或不出现。我们利用线性判别模型来设计 MMCRF 的势函数，其中边势函数定义为与图像视觉特征相关的平滑函数。我们设计了一个新颖的边缘损失函数来对 MMCRF 进行最大边缘参数估计。该损失函数把传统的 Hinge 损失分解为一组子 Hinge 损失的和，使得我们可以利用所提

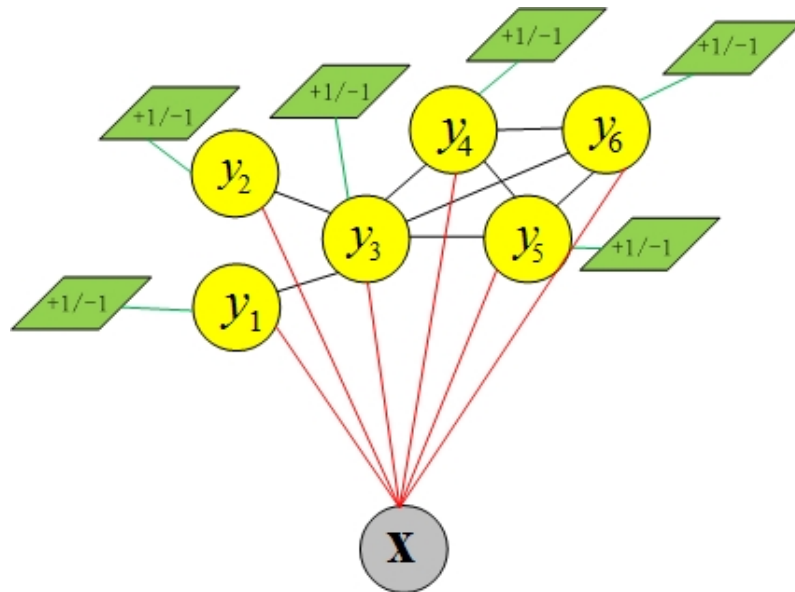


图 4.1 MMCRF 模型的基本框架

出的上下文核函数（Contextual Kernel）求解一系列独立的二次规划（QP）问题来估计 MMCRF 的参数。因此，MMCRF 对于语义空间具有可延展性。图 4.1 描绘了 MMCRF 模型的基本框架。从图中可以看出，MMCRF 把图像 \mathbf{x} 看作一个整体，属于基于场景的上下文相关模型。

本章余下内容组织如下：4.1 节介绍了 CRF 的基本概念；4.2 节描述了 MMCRF 图结构的构造；4.3 节给出了 MMCRF 势函数的设计；4.4 节和 4.5 节分别介绍了 MMCRF 的最大边缘参数估计和模型推理；4.6 节给出了 MMCRF 图像语义自动标注的实验结果；最后在 4.7 节对本章进行小结。

4.1. 条件随机场

条件随机场（CRF）是马尔科夫随机场的一个变形。CRF 与 MRF 不同之处在于，CRF 对随机变量的条件概率建模。令 $\mathbf{x} \in X$ 表示输入空间中的一个观察值， $\mathbf{y} \in Y$ 表示输出空间中的一个结果，其中 $\mathbf{y} = (y_1, y_2, \dots, y_m)^T$ 是一个随机向量。那么，给定观察值 \mathbf{x} ，随机变量集合 $\{y_1, y_2, \dots, y_m\}$ 被称作是点集 $S = \{1, 2, \dots, m\}$ 上，关于邻域系统 $N = \{N_i | i \in S\}$ 的条件随机场，其中 N_i 是与点 i 相邻的点，当且仅当满足下面两个条件：

$$(1) P(\mathbf{y} | \mathbf{x}) > 0, \forall \mathbf{y} \in Y,$$

$$(2) P(y_i | \mathbf{x}, y_{S-\{i\}}) = P(y_i | \mathbf{x}, y_{N_i}), \forall i \in S,$$

其中 $y_A = \{y_i | i \in A\}$ ， A 表示任意的点的集合。条件(2)是 CRF 中的马尔科夫性。根据 Hammersley-Clifford 定理[17]，每一个 CRF 都满足下面的条件概率密度分布：

$$P(\mathbf{y} | \mathbf{x}) = Z^{-1} \times e^{-U(\mathbf{x}, \mathbf{y})}, \quad (4.1)$$

其中

$$Z = \sum_{\mathbf{y} \in Y} e^{-U(\mathbf{x}, \mathbf{y})} \quad (4.2)$$

是划分函数， $U(\mathbf{x}, \mathbf{y})$ 是能量函数，它是所有集簇 $c \in C$ 上势函数 $V_c(\mathbf{x}, \mathbf{y})$ 的和：

$$U(\mathbf{x}, \mathbf{y}) = \sum_{c \in C} V_c(\mathbf{x}, \mathbf{y}).$$

当仅考虑点集大小不超过 2 的集簇时，能量函数可简化为：

$$U(\mathbf{x}, \mathbf{y}) = \sum_{i \in S} V_1(\mathbf{x}, y_i) + \sum_{i \in S} \sum_{j \in N_i} V_2(\mathbf{x}, y_i, y_j), \quad (4.3)$$

其中 $V_1(\mathbf{x}, y_i)$ 被称为点势函数， $V_2(\mathbf{x}, y_i, y_j)$ 被称为边势函数。

4.2. 概念图

我们根据训练数据集中语义概念的共同出现来构造 MMCRF 模型的图结构。给定

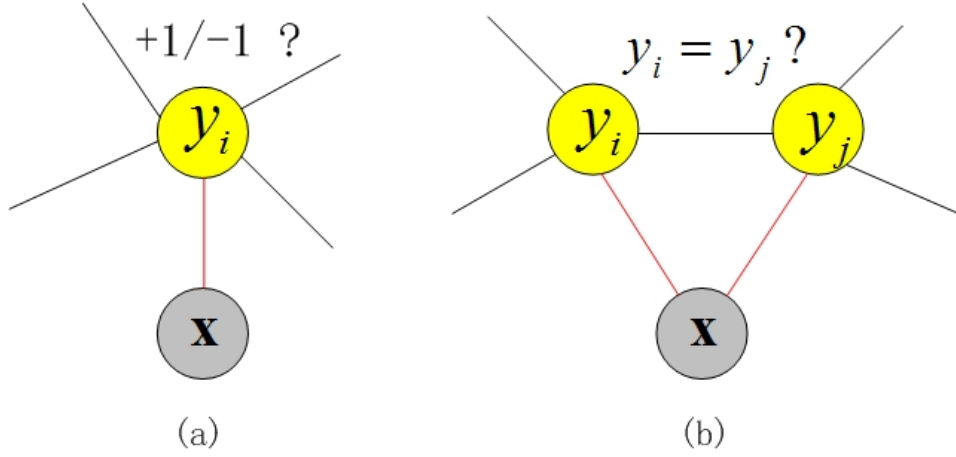


图 4.2 MMCRF 模型势函数的设计: (a)点势函数, (b)边势函数

训练集 $T = \{(\mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^T$, 其中观察值 \mathbf{x} 表示图像的视觉特征, 而输出 \mathbf{y} 表示语义概念的标签。两个语义概念共同出现当且仅当它们同时作为训练集中同一个观察值的标注词出现。我们基于语义概念的共现, 定义两个语义概念相关性的度量如下:

$$P(y_j | y_i) = \frac{|y_i \cap y_j|}{|y_i|}. \quad (4.4)$$

等式(4.4)的度量可看作是观察到 y_i 后 y_j 出现的先验条件概率的估计。利用上述相关性度量, 我们定义点 j 是点 i 的邻居, 即 $j \in N_i$, 当且仅当 $P(y_j | y_i) \geq P_0, \forall i, j \in S$, 其中 P_0 是预先定义的阈值常数。需要注意的是, 由于语义概念之间的相互作用不是对等的, 因此我们构造的邻域系统不是对称的。

4.3. 势函数设计

在 MMCRF 模型中, 点势函数 $V_1(\mathbf{x}, y_i)$ 根据观察值 \mathbf{x} 来判断标签 $y_i \in \{-1, +1\}$ 的值。图 4.2(a)描绘了点势函数设计的想法。由于线性判别模型具有坚实的理论基础, 并且模型的能力能够通过把输入映射到高维空间得到提升, 因此我们采用线性判别模型来对点势函数建模。MMCRF 的点势函数定义为:

$$V_1(\mathbf{x}, y_i) = -y_i (\mathbf{w}_i^T \phi_i(\mathbf{x}) + b_i), \quad (4.5)$$

其中 \mathbf{w}_i 和 b_i 是点 i 上的参数, ϕ_i 是将观察值 \mathbf{x} 映射到一个与语义概念 i 相关的空间中的函数。我们将在后面的章节中介绍 ϕ_i 的设计。在点势函数(4.5)中, $y_i(\mathbf{w}_i^T \phi_i(\mathbf{x}) + b_i)$ 是样本 $(\phi_i(\mathbf{x}), y_i)$ 相对于超平面 (\mathbf{w}_i, b_i) 的函数边缘 (functional margin)。因此, 增大点势函数线性模型的边缘将会降低势能。

我们同样采用线性判别模型来对 MMCRF 的边势函数 $V_2(\mathbf{x}, y_i, y_j)$ 建模。与点势函数线性模型不同的是, 边势函数线性模型根据观察值 \mathbf{x} 来判断两个点上的一对标签应

该相同还是不同。图 4.2(b)描绘了 MMCRF 边势函数的设计。边势函数可以看作一个与视觉特征相关的平滑函数。令边势函数线性模型中的截距为 0, 边势函数可定义为:

$$V_2(\mathbf{x}, y_i, y_j) = -y_i y_j \mathbf{w}_{ij}^T \phi_j(\mathbf{x}), \quad (4.6)$$

其中 \mathbf{w}_{ij} 是边 (i, j) 上的参数, ϕ_j 是将观察值 \mathbf{x} 映射到一个与语义概念 j 相关的空间中的函数。边上的参数 \mathbf{w}_{ij} 并不是对称的。在等式(4.6)中, 如果我们把 $y_i y_j$ 看作是样本 $\phi_j(\mathbf{x})$ 在边势函数线性模型中的标签, 那么增大边势函数线性模型的边缘也等价于减小势能。将等式(4.5)和(4.6)代入等式(4.3)中, 我们得到 MMCRF 的能量函数:

$$U(\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{b}) = -\sum_{i \in S} y_i (\mathbf{w}_i^T \phi_i(\mathbf{x}) + b_i) - \sum_{i \in S} \sum_{j \in N_i} y_i y_j \mathbf{w}_{ij}^T \phi_j(\mathbf{x}), \quad (4.7)$$

其中 $\mathbf{w} = \{\mathbf{w}_i, \mathbf{w}_{ij} \mid \forall i \in S, j \in N_i\}$, $\mathbf{b} = \{b_i \mid \forall i \in S\}$ 表示 MMCRF 的参数。

4.4. 最大边缘参数估计

在基于能量的学习框架下[40], 我们不需要对 MMCRF 进行正规化。参数估计的任务是为了寻找到某一特定的能量函数, 使得该能量函数对于给定观察值的最小值对应的标签是正确标签。我们通过定义损失函数来评价不同的能量函数。所定义的损失函数涵盖了训练数据集上的损失和我们对于当前任务的先验知识。从而, 参数估计转化为寻找使得损失函数产生最小值的参数。

4.4.1. 拆分的 Hinge 损失

边缘损失函数会在正确的标签与错误的标签之间构造出能量差。边缘大的分类器将会具有好的泛化性能。Hinge 损失是一种边缘损失函数, 它被应用于 SVM [41]和 M3N [37]。我们利用 Hinge 损失来估计 MMCRF 模型的参数。一个训练样本 $(\mathbf{x}^t, \mathbf{y}^t)$ 上的 Hinge 损失定义为:

$$L_{\text{Hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) = \max(0, m + U(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) - U(\mathbf{x}^t, \bar{\mathbf{y}}^t, \mathbf{w}, \mathbf{b})), \quad (4.8)$$

其中 $m > 0$ 是正的边缘, $\bar{\mathbf{y}}^t$ 是最违反的错误标签:

$$\bar{\mathbf{y}}^t = \arg \min_{\mathbf{y} \in Y \text{ 且 } \mathbf{y} \neq \mathbf{y}^t} U(\mathbf{x}^t, \mathbf{y}, \mathbf{w}, \mathbf{b}). \quad (4.9)$$

当正确标签与最违反的错误标签之间的能量差大于 $-m$ 时, Hinge 损失(4.8)与该能量差成正比。将等式(4.7)代入等式(4.8), 并重新组合各项和式可得:

$$L_{\text{Hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) = \max \left(0, m + \sum_{i \in S} (\bar{y}_i^t - y_i^t) (\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i) + \sum_{i \in S} \sum_{j \in N_i} (\bar{y}_i^t \bar{y}_j^t - y_i^t y_j^t) \mathbf{w}_{ij}^T \phi_j(\mathbf{x}^t) \right). \quad (4.9)$$

在上述的 Hinge 损失(4.9)中, 最违反的错误标签是未知的。因此, 为了保证 Hinge 损失的边缘达到 m , 我们需要保证输出空间 Y 中的每一个错误标签与正确标签的能量差都达到 m 。这样将会导致需要求解一个约束数目为语义概念数目指数倍的最优化问题。为了解决这一问题, 结构化 SVM [13] 维护一个激活约束的工作集, 在迭代过程中仅仅考虑该工作集中的约束。而 M3N [37] 利用标签间的马尔科夫性将最优化问题的约束数目降低到多项式级。但是, 结构化 SVM 和 M3N 最终都需要求解一个较为复杂的最优化问题。与先前的工作不同, 我们基于 MMCRF 势函数的设计, 提出一种拆分的 Hinge 损失来进行参数估计。在拆分的 Hinge 损失中, 点势函数线性模型和边势函数线性模型的最违反的错误标签就是相应的正确标签取反:

$$L'_{\text{Hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) = \sum_{i \in S} \max(0, m_i - 2y_i^t(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i)) + \sum_{i \in S} \sum_{j \in N_i} \max(0, m_{ij} - 2y_i^t y_j^t \mathbf{w}_{ij}^T \phi_j(\mathbf{x}^t)), \quad (4.10)$$

其中 $m_i, i \in S$ 是点势函数线性模型的边缘, $m_{ij}, i \in S, j \in N_i$ 是边势函数线性模型的边缘。下面的命题表明, 我们可以用拆分的 Hinge 损失(4.10)来代替 Hinge 损失(4.9)进行参数估计。

命题 4.1 如果 $m \leq \sum_{i \in S, \bar{y}_i^t = -y_i^t} m_i + \sum_{i \in S, j \in N_i, \bar{y}_i^t \bar{y}_j^t = -y_i^t y_j^t} m_{ij}$, 那么 $L'_{\text{Hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b})$ 是

$L_{\text{Hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b})$ 的上界。

根据命题 4.1, 如果我们令 m 小于某个阈值, 那么 L'_{Hinge} 将是 L_{Hinge} 的上界, 减小拆分的 Hinge 损失 L'_{Hinge} 也会同时减小原来的 Hinge 损失 L_{Hinge} 。采用 L'_{Hinge} 进行参数估计, 我们可以分别求解每一个线性模型的最大边缘超平面。但是, 在模型推理时, 点势函数线性模型与边势函数线性模型之间存在紧密的相互作用。它们之间的错误传播会降低模型的性能。为了减小错误传播的影响, 我们采用了 L_{Hinge} 与 L'_{Hinge} 之间的一个折衷。我们把 L_{Hinge} 拆分到点, 使得 MMCRF 的参数估计可以逐点进行。每一个点势函数线性模型 (\mathbf{w}_i, b_i) 和与之相邻的边势函数线性模型 $\mathbf{w}_{ij}, j \in N_i$ 同时训练出来。相应的拆分的 Hinge 损失定义为:

$$L''_{\text{Hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) = \sum_{i \in S} \max \left(0, m_i - 2y_i^t(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i) - \sum_{j \in N_i} 2y_i^t y_j^t \mathbf{w}_{ij}^T \phi_j(\mathbf{x}^t) \right). \quad (4.11)$$

在 L''_{Hinge} 中, 边势函数线性模型的边缘并没有明确地表示出来。因此, m 的上界并不能像命题 4.1 一样表示出来。但是 m 上界的存在能够保证在 m 足够小时, L''_{Hinge} 是 L_{Hinge} 的上界。从而我们可以采用 L''_{Hinge} 来估计 MMCRF 的参数。

4.4.2. 有偏向的正则化

我们根据图像语义自动标注任务的先验知识，在损失函数中加入正则化项来避免模型过拟合。由于 CRF 边上控制标签相互作用的参数通过会被训练为起到过高的作用 [36]，我们需要对边势函数线性模型进行更多的惩罚。因此，我们在正则化项中引入两个不同的参数分别控制对点势函数线性模型和边势函数线性模型的惩罚。有偏向的正则化项定义为：

$$R(\mathbf{w}) = \lambda_1 \sum_{i \in S} \|\mathbf{w}_i\|^2 + \lambda_2 \sum_{i \in S} \sum_{j \in N_i} \|\mathbf{w}_{ij}\|^2, \quad (4.12)$$

其中 λ_1 和 λ_2 是两个不同的参数，它们分别控制对点势函数线性模型和边势函数线性模型的惩罚。

4.4.3. 参数估计框架

给定训练数据集 $T = \{(\mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^{|T|}$ ，我们将单个样本上拆分的 Hinge 损失(4.11)和有偏向的正则化项(4.12)组合起来得到下面的损失函数：

$$L(T, \mathbf{w}, \mathbf{b}) = \frac{1}{|T|} \sum_{t=1}^{|T|} L_{\text{Hinge}}''(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) + R(\mathbf{w}). \quad (4.13)$$

采用与支持向量机[41]相同的方法，我们将损失函数乘以 1/2，令点势函数线性模型的边缘为 1，再引入松弛变量允许边缘的违反，则可以得到以下参数估计最优化问题的主要形式：

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{b}, \xi} \quad & \frac{1}{2} \left(\sum_{i \in S} \|\mathbf{w}_i\|^2 + \lambda \sum_{i \in S} \sum_{j \in N_i} \|\mathbf{w}_{ij}\|^2 \right) + C \sum_{i \in S} \sum_{t=1}^{|T|} \xi_i^t \\ \text{s.t.} \quad & y_i^t \left(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i + \sum_{j \in N_i} y_j^t \mathbf{w}_{ij}^T \phi_j(\mathbf{x}^t) \right) \geq 1 - \xi_i^t, \\ & \xi_i^t \geq 0, \forall i \in S, \forall t = 1, \dots, |T| \end{aligned} \quad (4.14)$$

其中 $C = \frac{1}{\lambda_1 |T|}$ 和 $\lambda = \frac{\lambda_2}{\lambda_1}$ 是两个常数， $\xi = \{\xi_i^t \mid \forall i \in S, \forall t = 1, \dots, |T|\}$ 表示引入的松弛变量。

4.4.4. 利用上下文核函数求解最优化问题的算法

MMCRF 模型的最大边缘参数估计转化为求解最优化问题(4.14)。仔细观察等式

(4.14)可知, 该最优化问题可分解为 $|S|$ 个子问题, 分别对应 CRF 的每个点。因此, 我们可以分别估计 MMCRF 每个点的参数。点 i 对应的最优化子问题为:

$$\begin{aligned} \min_{\mathbf{w}_{i \cup N_i}, b_i, \xi_i} & \frac{1}{2} \left(\|\mathbf{w}_i\|^2 + \lambda \sum_{j \in N_i} \|\mathbf{w}_{ij}\|^2 \right) + C \sum_{t=1}^{|T_i|} \xi_i^t \\ \text{s.t. } & y_i^t \left(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i + \sum_{j \in N_i} y_j^t \mathbf{w}_{ij}^T \phi_j(\mathbf{x}^t) \right) \geq 1 - \xi_i^t, \\ & \xi_i^t \geq 0, \forall t = 1, \dots, |T_i| \end{aligned} \quad (4.15)$$

其中 $\mathbf{w}_{i \cup N_i} = \{\mathbf{w}_i, \mathbf{w}_{ij} \mid j \in N_i\}$, $\xi_i = \{\xi_i^t \mid \forall t = 1, \dots, |T_i|\}$ 。在等式(4.15)中, 我们从整体训练数据集 T 中采样为点 i 构造了专用的训练集 $T_i = \{(\mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^{|T_i|}$, 使得 T_i 相对于语义概念 i 具有更为平衡的正负样例。经过拉格朗日变换后, 我们得到最优化问题(4.15)的对偶形式:

$$\begin{aligned} \max_{\alpha_i} & -\frac{1}{2} \sum_{t=1}^{|T_i|} \sum_{t'=1}^{|T_i|} \alpha_i^t \alpha_i^{t'} y_i^t y_i^{t'} K_i(\mathbf{x}^t, \mathbf{x}^{t'}) + \sum_{t=1}^{|T_i|} \alpha_i^t \\ & - \frac{1}{2\lambda} \sum_{t=1}^{|T_i|} \sum_{t'=1}^{|T_i|} \sum_{j \in N_i} \alpha_i^t \alpha_i^{t'} y_i^t y_j^{t'} y_j^{t'} K_j(\mathbf{x}^t, \mathbf{x}^{t'}), \\ \text{s.t. } & \sum_{t=1}^{|T_i|} \alpha_i^t y_i^t = 0, C \geq \alpha_i^t \geq 0, \forall t = 1, \dots, |T_i| \end{aligned} \quad (4.16)$$

其中 $\alpha_i = \{\alpha_i^t \mid \forall t = 1, \dots, |T_i|\}$ 表示对偶变量。与 SVM 相同, 我们用核函数 $K_i(\mathbf{x}^t, \mathbf{x}^{t'})$ 和 $K_j(\mathbf{x}^t, \mathbf{x}^{t'})$ 分别替代了观察值在映射后的特征空间中的内积 $\langle \phi_i(\mathbf{x}^t) \cdot \phi_i(\mathbf{x}^{t'}) \rangle$ 和 $\langle \phi_j(\mathbf{x}^t) \cdot \phi_j(\mathbf{x}^{t'}) \rangle$ 。因此, 我们不需要专门设计与语义概念相关的映射 ϕ , 而只需要为每一个语义概念构造自身的核函数。仔细观察对偶问题(4.16)可知, 它是具有以下核函数的最大边缘分类器的对偶形式:

$$K(\mathbf{x}^t, \mathbf{x}^{t'}) = K_i(\mathbf{x}^t, \mathbf{x}^{t'}) + \frac{1}{\lambda} \sum_{j \in N_i} y_j^t y_j^{t'} K_j(\mathbf{x}^t, \mathbf{x}^{t'}). \quad (4.17)$$

我们推导出的核函数(4.17)不仅利用了图像的视觉特征, 而且利用了点 i 邻域内点上的标签。因此, 我们把等式(4.17)称为上下文核函数 (Contextual Kernel)。上下文核函数中后面一项和式可以看作是“平滑”核函数。我们通过参数 λ 来控制它在上下文核函数中的比重, 从而保证整个核函数是半正定的。通过上下文核函数, 我们可以利用普通的 SVM 算法来求解对偶问题(4.16)。

4.4.5. 核函数的构造

不同种类的视觉特征, 例如颜色, 纹理, 局部外观, 等等, 对于识别不同语义概念的贡献并不相同。因此, 我们采用判别度量学习将图像不同种类的视觉特征组合起

来构造面向语义的核函数。我们首先对图像提取出多种视觉特征。其次，我们在每一种特征上计算两幅图像之间的基本距离。最后，两幅图像之间的距离定义为它们之间各种特征上基本距离的加权和：

$$d_{\hat{\mathbf{w}}}(\mathbf{x}^t, \mathbf{x}^{t'}) = \hat{\mathbf{w}}^T \mathbf{d}_{tt'}, \quad (4.18)$$

其中 $\mathbf{d}_{tt'}$ 是由 \mathbf{x}^t 与 $\mathbf{x}^{t'}$ 之间基本距离组成的距离向量， $\hat{\mathbf{w}}$ 是需要从训练集里学习的距离权重。在加权距离的基础上，我们采用泛化的高斯核作为面向语义的核函数：

$$K_{\hat{\mathbf{w}}}(\mathbf{x}^t, \mathbf{x}^{t'}) = e^{-g d_{\hat{\mathbf{w}}}(\mathbf{x}^t, \mathbf{x}^{t'})}, \quad (4.19)$$

其中 g 是高斯核的宽度。我们采用基于最近邻模型的判别度量学习算法[22]来学习核函数中的权重 $\hat{\mathbf{w}}$ 。不同之处在于，学习点 i 上的核函数 K_i 时，我们仅仅采用训练集中语义概念 i 的正样例，从而得到了不同的面向语义的核函数。

4.5. 模型推理

CRF 中模型推理的任务是找到与给定观察值最兼容的标签组合。形式上讲，给定观察值 \mathbf{x} ，推理产生具有最小能量的标签组合：

$$\mathbf{y}^* = \arg \min_{\mathbf{y} \in Y} U(\mathbf{x}, \mathbf{y}), \quad (4.20)$$

其中 $U(\mathbf{x}, \mathbf{y})$ 在等式(4.3)中定义。由于标签组合的数目是语义概念数目的指数倍，枚举搜索是行不通的。我们采用迭代条件模式 (Iterated Conditional Modes, ICM) [38] 算法来进行模型推理。ICM 通过最大化局部条件概率来顺序更新每一个标签。这等价于最小化下面的每个点上的局部能量函数：

$$\begin{aligned} U_i(\mathbf{x}, y_i, y_{N_i}) &= V_1(\mathbf{x}, y_i) + \sum_{j \in N_i} V_2(\mathbf{x}, y_i, y_j) \\ &= -y_i \left(\mathbf{w}_i^T \phi_i(\mathbf{x}) + b_i \right) - \sum_{j \in N_i} y_i y_j \mathbf{w}_{ij}^T \phi_j(\mathbf{x}). \end{aligned} \quad (4.21)$$

在第 $(t+1)$ 步迭代中，给定观察值 \mathbf{x} 和点 i 邻域内的标签 $y_{N_i}^{(t)}$ ，ICM 算法利用以下规则把每一个 $y_i^{(t)}$ 更新为 $y_i^{(t+1)}$ ：

$$y_i^{(t+1)} = \arg \min_{y_i} U_i(\mathbf{x}, y_i, y_{N_i}^{(t)}). \quad (4.22)$$

上述更新规则等价于

$$y_i^{(t+1)} = \begin{cases} +1, & \text{if } \mathbf{w}_i^T \phi_i(\mathbf{x}) + b_i + \sum_{j \in N_i} y_j^{(t)} \mathbf{w}_{ij}^T \phi_j(\mathbf{x}) \geq 0 \\ -1, & \text{otherwise.} \end{cases} \quad (4.23)$$

我们采用对偶变量和核函数来表述以上更新规则：

$$y_i^{(t+1)} = \begin{cases} +1, & \text{if } \sum_{t=1}^{|T_i|} \alpha_i' y_i' K_i(\mathbf{x}', \mathbf{x}) + b_i + \frac{1}{\lambda} \sum_{t=1}^{|T_i|} \sum_{j \in N_i} \alpha_i' y_i' y_j^{(t)} K_j(\mathbf{x}', \mathbf{x}) \geq 0 \\ -1, & \text{otherwise,} \end{cases} \quad (4.24)$$

其中 α_i 和 b_i 是估计好的参数。从一个初始标签组合开始（所有标签设置为 -1 ），ICM 迭代到算法收敛。我们则得到了与观察值 \mathbf{x} 近似最兼容的标签组合。

4.6. 实验

4.6.1. 实验数据集

Corel 数据集：我们采用 Corel 数据集[8]来进行实验。该数据集被广泛应用于图像语义自动标注的性能比较。Corel 数据集包含 5000 幅图像，其中 4500 幅是训练图像，其余 500 幅图像用于测试。每幅图像被标注了 1-5 个语义关键词，而数据集中一共有 374 个关键词。大部分的关键词都只具有少量的正样例。例如，374 个关键词中只有 70 个关键词的正样例数目超过 60 个。

TRECVID-2005 数据集：该数据集包含有 108 小时多种语言的广播新闻视频。这些视频被切割为 61901 帧。视频关键帧的内容与 Corel 图像相比更为丰富且代表现实世界中的具体场景。数据集中包含有 39 个语义概念，每帧图像被标注上 0-11 个语义概念。在整个数据集上实验是相当花费时间的。因此，我们分别从数据集中的 90 个视频和其余的 47 个视频中选择训练数据和测试数据。对于每个语义概念，我们分别随机选取不超过 500 和 100 个正样例作为训练和测试数据。最终我们有 6657 帧图像用于训练和 1748 帧图像用于测试。

4.6.2. 特征提取

我们从数据集中提取图像检索与分类中常用的多种视觉特征。我们采用两种全局特征：Gist 特征[27]和颜色直方图。颜色直方图的计算是分别在颜色的三个通道中采用八个箱来计算。我们采用了颜色的三种表示形式：RGB，LAB 和 HSV，结果每幅图像得到了三个 512 维的颜色特征向量。对于局部特征，我们采用 SIFT[42]，并且采用了“Bag-of-features”的软加权模式[43]。除 Gist 特征外的所有特征向量都经过了 L1 正规化。为了计算同种特征之间的基本距离，我们对于 Gist 采用 L2 度量，对于颜色特征采用 L1 度量，而对于 SIFT 采用 χ^2 度量。

4.6.3. 评价度量

我们采用查全率 (recall), 查准率 (precision) 和 F1 来度量标注性能。给定一个查询关键词 w , 令 $|W_G|$ 表示测试集中人工标注为 w 的图像的数目, $|W_M|$ 表示标注算法标注为 w 的测试图像的数目, $|W_C|$ 表示标注算法标注正确的测试图像的数目, 则查全率, 查准率与 F1 分别定义为:

$$recall = \frac{|W_C|}{|W_G|}, \quad precision = \frac{|W_C|}{|W_M|}, \quad F1 = \frac{2 \times recall \times precision}{recall + precision}.$$

我们分别对每个关键词计算查全率和查准率, 再对每个关键词的查全率和查准率求平均作为标注性能的评价度量。

4.6.4. 语义上下文建模评价

我们选择 SVM 作为基准方法, 通过 MMCRF 与 SVM 在 Corel 数据集上的对比来评价 MMCRF 的语义上下文建模。我们为 Corel 数据集中的每一个关键词训练一个二分类的 SVM。由于现实世界中的数据通常是不平衡的, 因此为每一个关键词构造自身的训练数据集是相当重要的。为了捕获不同关键词的语义, 我们采用了每个关键词的全部正样例。当关键词的正样例数目较少时, 采用相同数目的负样例会导致模型预测过多的错误正样例。因此, 对于这样的关键词, 我们选取更多的负样例。为了说明语义上下文的作用并且得到一个公平的比较, 我们对于 MMCRF 和 SVM 采用相同的负样例选择策略。每个关键词至少使用了 200 个负样例。两种模型都采用了从多种图像视觉特征中利用判别度量学习得到的核函数。表 4.1 列出了这两种模型对比的实验结果。我们对于 MMCRF 中参数 λ 的不同取值进行了实验。参数 λ 可以控制语义概念之间相互作用的大小。 λ 的取值越大, 语义上下文起到的作用反而越小。表 4.1 中仅仅列出了一些具有代表性的结果。当 $\lambda=1$ 时, MMCRF 预测出了测试集 263 个关键词中的 146 个关键词, 远远超过 SVM 的 81 个。但是 MMCRF 的平均标注长度超过了 30。这说明语义概念的相互作用过于强烈, 使得模型的性能降低。当 $\lambda=140$ 时, 我们的方法的标注性能与 SVM 相差不大。这说明此时语义上下文对于标注性能的影响不大。当 λ 的取值在这两个极值之间, 并且模型的平均标注长度约等于 5 时, MMCRF 预测出了所有 263 个关键词中的 99 个, 预测出了出现频率最高的 70 个关键词中的 68 个。这两个数据都大于 SVM 的 81 个和 63 个。当 $\lambda=80$ 时, MMCRF 在出现频率最高的 70 个关键词上的平均查全率和平均查准率分别比 SVM 提高了 2% 和 10%。实验结果表明, 通过对语义上下文建模, 我们的方法能够提高出现频率高的关键词的标注性能, 并且有能力预测出出现频率较低的关键词。

表 4.1 MMCRF 与 SVM 在 Corel 数据集上的性能对比。N+, Length, R 和 P 分别表示查全率大于 0 的关键词数目, 平均标注长度, 平均查全率和平均查准率。263 和 70 分别表示出现在测试集中的 263 个关键词和数据集中出现频率最高的 70 个关键词。

模型	SVM	$\lambda = 1$	$\lambda = 60$	$\lambda = 80$	$\lambda = 140$
N+ of 263	81	146	97	99	87
Length	4.33	33.85	5.15	4.97	4.90
R of 70	0.5447	0.4725	0.5226	0.5554	0.5393
P of 70	0.3983	0.1450	0.4409	0.4373	0.3982
N+ of 70	63	63	68	67	64

4.6.5. 在 Corel 数据集上的对比

为了更进一步地评估我们提出的模型的性能, 我们在 Corel 数据集上将其与多种当前最新的标注方法进行对比, 其中包括四种上下文无关模型: SVM, ASVM-MIL [20], MBRM [3]和 TagProp [22], 两种上下文相关模型: HCM[11]和 MMRF[15]。图 4.3 给出了我们的方法与四种上下文无关模型的实验结果。因为判别模型不能有效地捕获正样例数目较少的关键词的语义[20], 我们仅仅在出现频率最高的 70 个关键词上比较不同模型的性能。对于 SVM, TagProp 和 MMCRF, 我们在五种视觉特征上利用判别度量学习算法学习不同特征的权重。对于 SVM 和 MMCRF, 我们采用交叉检验来确定每个关键词的训练时采用的负样例数目。从图 4.3 中可以看出, 我们的模型取得了最

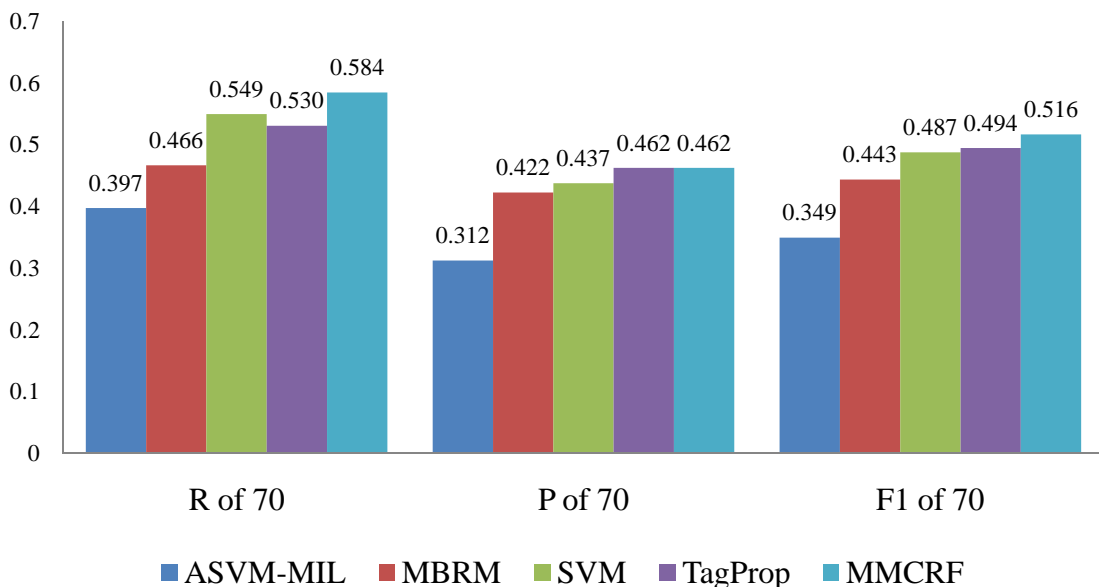


图 4.3 MMCRF 与四种上下文无关模型在 Corel 数据集上的性能比较

表 4.2 MMCRF 与两种上下文相关模型在 Corel 数据集上的性能比较

模型	MMRF	MMCRF	HCM	MMRF	MMCRF
	70 个关键词上的结果		104 个关键词上的结果		
N+	69	70	87	97	88
R	0.518	0.584	0.433	0.427	0.444
P	0.448	0.462	0.359	0.445	0.402
F1	0.480	0.516	0.393	0.437	0.422

高的 F1。与性能第二高的方法：TagProp 相比，MMCRF 对于 70 个关键词上的平均查全率取得了 10% 的改进，同时取得了与 TagProp 相同的平均查准率。

表 4.2 列出了 MMCRF 与两种上下文相关模型：HCM 和 MMRF 在 Corel 数据集上的性能比较。我们的方法在出现频率最高的 70 个关键词上的平均查全率和平均查准率分别比 MMRF 改进了 13% 和 3%。我们从[11]中得到了 HCM 在出现频率最高的 104 个关键词的实验结果。因此，我们也在这 104 个关键词上计算了 MMCRF 和 MMRF 的实验结果。从表 4.2 中可知，MMCRF 在出现频率最高的 104 个关键词上的平均查全率和平均查准率分别比 HCM 改进了 3% 和 12%。因为 MMCRF 采取了最大边缘参数估计的框架，所以 MMCRF 与 MMRF 相比，对于关键词的正样例数目更为敏感。而第 71 到第 104 个出现频率最高的关键词的正样例数目较少，这导致了 MMCRF 在这些关键词上标注性能的降低。因此，MMRF 在出现频率最高的 104 个关键词上取得了更高的 F1。我们在表 4.3 中列出了 MMCRF 在 Corel 数据集上的一些标注样例，同时给出了相应的人工标注。我们选择了不同场景的图像。从标注结果可以看出，我们的方法能够抓住图像内容的主旨。

4.6.6. 在 TRECVID-2005 数据集上的对比

我们在 TRECVID-2005 数据上测试 MMCRF 对于视频标注的性能。由于 SVM 被广泛应用于视频中语义概念的检测，而 TagProp [22]在 Corel 数据集上取得了很有竞争力的标注性能，因此我们在 TRECVID-2005 数据集上与上述两种方法进行对比。图 4.4 给出了对比试验的结果。从图中可以看出，MMCRF 在 39 个语义概念上的平均查全率和平均查准率均比 SVM 高出了 3%。这一改进并没有 Corel 数据集上的改进显著，是因为 TRECVID-2005 数据集的语义空间较小，限制了语义上下文的作用。值得注意的是，MMCRF 和 SVM 在 TRECVID-2005 数据集上的性能均超过了 TagProp。MMCRF 相对于 TagProp 在 39 个语义概念上的平均查全率和平均查准率分别提高了 11% 和 4%。这一结果说明，基于最近邻模型的方法，例如 TagProp，对于图像内容多样性大的数

表 4.3 MMCRF 标注结果与人工标注在 Corel 和 TRECVID-2005 数据集上的对比

Corel					
人工标注	branch bird nest	wall car track formula	building clothes shop street	stone statue sculpture sphinx	tree snow wood fox
MMCRF	tree branch grass bird nest	wall car track	people building shop street	stone statue sculpture	snow rock fox
TRECVID-2005					
人工标注	Face Flag-US Person Government -leader	Face Map Person Studio	Outdoor People-Marching	Animal Mountain Sky Outdoor Vegetation Waterscape_Water front	Airplane Outdoor Sky
MMCRF	Face Flag-US Person Meeting Government -leader	Face Map Person Studio	Crowd Outdoor People-Marching Person	Animal Boat_Ship Mountain Outdoor Sky Waterscape_Water front	Airplane Outdoor Sky

数据集性能会下降。而判别模型，例如 MMCRF 和 SVM 具有很强的能力来处理现实世

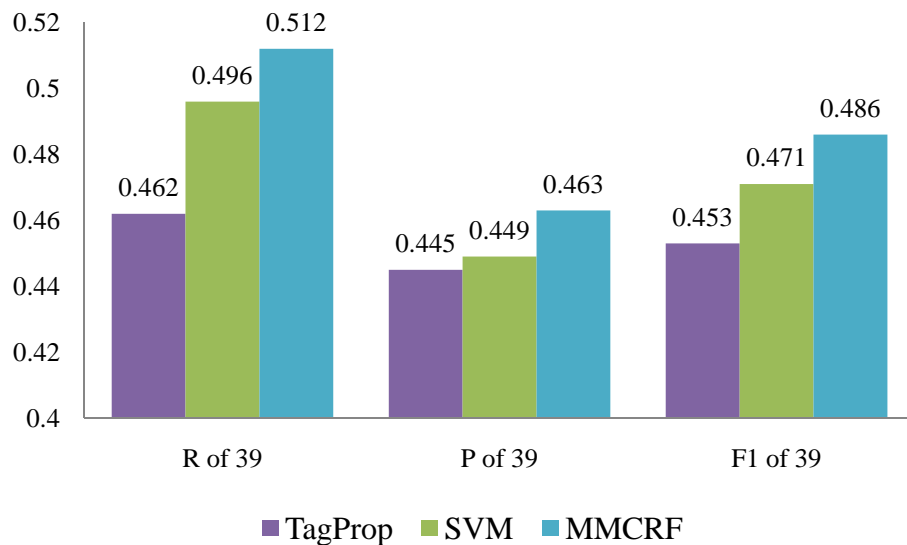


图 4.4 MMCRF 与 TagProp 和 SVM 在 TRECVID-2005 数据集上的性能比较

界数据集中的内容多样性。表 4.3 给出了 TRECVID-2005 数据集中 MMCRF 的一些标注结果，同时列出了相应的人工标注。从图中可以看出，MMCRF 取得了令人满意的标注。例如，对于第二和第五帧图像，MMCRF 的标注与人工标注完全一致。MMCRF 甚至标注出了人工忽略的正确标注，例如第三帧图像中的“Crowd”。实验结果表明，MMCRF 具有处理不同场景图像标注的能力。

4.7. 本章小结

我们在本章中提出了一个新颖的最大边缘条件随机场上下文相关模型对图像语义自动标注中的语义上下文建模。我们的方法继承了最大边缘学习方法的优点，同时能够准确地捕获到语义概念之间的相关性。我们从判别分析的角度设计了 MMCRF 的势函数，并提出了一个新颖的拆分的 Hinge 损失对 MMCRF 进行最大边缘参数估计。我们利用推导出的上下文核函数和普通的 SVM 算法求解一系列 QP 问题来训练 MMCRF。在图像和视频数据上的大量实验结果表明，我们的方法能够利用语义概念之间的相关性来处理各种场景的图像语义自动标注问题。

第五章 结束语

5.1. 本文贡献

本文提出了一种新颖的采用马尔科夫随机场对图像语义自动标注中语义上下文建模的框架。通过对语义上下文建模，我们利用语义概念之间的相关性来提高语义自动标注的性能。在该 **MRF** 标注框架中，我们提出了两个新颖的语义上下文相关模型。一个是多马尔科夫随机场上下文相关模型，该模型在图像语义自动标注中生成模型的基础上为每一个语义概念构造语义层的 **MRF** 模型。另一个是最大边缘条件随机场上下文相关模型，该模型从判别分析的角度出发，利用条件随机场对语义上下文建模。

5.1.1. 马尔科夫随机场标注框架

本文工作的一个主要贡献是将马尔科夫随机场应用语义上下文建模。**MRF** 理论在计算视觉领域有着广泛的应用。先前的工作主要利用 **MRF** 对图像像素或区域的空间位置关系建模，即 **MRF** 中的每一个点表示一个图像像素或一块图像区域。本文提出了利用 **MRF** 对语义概念之间的相关性建模。在我们的 **MRF** 标注框架中，**MRF** 中的每一个点表示一个语义概念，两个点之间的边表示相应的两个语义概念相关。我们根据训练数据集中语义概念共同出现的频率来构造 **MRF** 的图结构。**MRF** 的每一个点上用一个随机变量来表示该点的标签。标签的取值为+1 或-1，分别表示相应的语义概念在给定的图像中出现或不出现。在 **MRF** 标注框架下构造语义上下文相关模型需要完成以下三步工作：势函数设计，参数估计和模型推理，其中势函数设计是模型构造的关键。我们可以在势函数中引入对于当前任务的先验知识。

5.1.2. 多马尔科夫随机场上下文相关模型

基于本文提出的 **MRF** 标注框架，我们提出了多马尔科夫随机场 (**MMRF**) 上下文相关模型。**MMRF** 是一个二层的模型，底层是图像语义自动标注中任一生成模型，用于估计图像的视觉特征与语义关键词共同出现的联合概率。高层是利用生成模型估计的联合概率构造的 **MRF** 模型。为了准确地把握不同关键词的语义，**MMRF** 对于每一个关键词构造自身的语义上下文模型。因此，我们可以把 **MMRF** 看作是利用语义概念之间的相关性对生成模型的标注结果进行改进。我们通过 **MRF** 引入了适当的参数设置来对语义上下文建模，大大增强了模型的学习能力。在 Corel 数据集和

TRECVID-2005 数据集上的大量实验结果表明, MMRF 能够利用语义概念之间的相关性, 显著地改进生成模型的标注结果。

5.1.3. 最大边缘条件随机场上下文相关模型

本文的另一个重要贡献是提出了最大边缘条件随机场(MMCRF)上下文相关模型。在考虑到 MMRF 模型的分层会导致模型的效率降低, 且模型的标注性能受到底层生成模型的限制, 我们在 MRF 标注框架下提出了 MMCRF 模型从判别分析的角度对语义上下文建模。MMCRF 把图像语义自动标注看作多标签分类问题, 在标注过程中利用语义概念之间的相关性来提高标注性能。我们采用线性判别模型来设计 MMCRF 的势函数, 其中边势函数可以看作是与图像视觉特征相关的平滑函数。我们提出了拆分的 Hinge 损失在最大边缘框架下估计 MMCRF 的参数, 使得模型相对于语义空间具有可扩展性。我们在 Corel 图像数据集和 TRECVID-2005 视频数据集上进行了实验。实验结果表明 MMCRF 能够利用语义概念之间的相关性改进各种场景的图像语义自动标注性能。

5.2. 将来工作

我们已经提出了采用马尔科夫随机场对图像语义自动标注中语义上下文建模的框架, 并且取得了令人鼓舞的实验结果。在将来的工作中, 我们将进一步探索图模型在语义上下文建模中的应用。我们将关注于提高模型对于语义概念间相关性的利用率。一个方向是探索 MRF 的图结构对于模型性能的影响, 另一个方向是采用更为高效的模型推理算法来改进模型的性能, 例如置信传播 (Belief Propagation) 算法。

附录

1. 命题 4.1 的证明

命题 4.1 如果 $m \leq \sum_{i \in S, \bar{y}_i' = -y_i'} m_i + \sum_{i \in S, j \in N_i, \bar{y}_i' \bar{y}_j' = -y_i' y_j'} m_{ij}$, 那么 $L'_{\text{Hinge}}(\mathbf{x}', \mathbf{y}', \mathbf{w}, \mathbf{b})$ 是 $L_{\text{Hinge}}(\mathbf{x}', \mathbf{y}', \mathbf{w}, \mathbf{b})$ 的上界。

证明：令

$$\begin{aligned} D &= L'_{\text{Hinge}}(\mathbf{x}', \mathbf{y}', \mathbf{w}, \mathbf{b}) - L_{\text{Hinge}}(\mathbf{x}', \mathbf{y}', \mathbf{w}, \mathbf{b}) \\ &= \sum_{i \in S} \max(0, m_i - 2y_i'(\mathbf{w}_i^T \phi_i(\mathbf{x}') + b_i)) + \sum_{i \in S} \sum_{j \in N_i} \max(0, m_{ij} - 2y_i' y_j' \mathbf{w}_{ij}^T \phi_j(\mathbf{x}')) \\ &\quad - \max\left(0, m + \sum_{i \in S} (\bar{y}_i' - y_i')(\mathbf{w}_i^T \phi_i(\mathbf{x}') + b_i) + \sum_{i \in S} \sum_{j \in N_i} (\bar{y}_i' \bar{y}_j' - y_i' y_j') \mathbf{w}_{ij}^T \phi_j(\mathbf{x}')\right) \end{aligned}$$

分以下两种情况讨论，

1) 如果

$$m + \sum_{i \in S} (\bar{y}_i' - y_i')(\mathbf{w}_i^T \phi_i(\mathbf{x}') + b_i) + \sum_{i \in S} \sum_{j \in N_i} (\bar{y}_i' \bar{y}_j' - y_i' y_j') \mathbf{w}_{ij}^T \phi_j(\mathbf{x}') \leq 0,$$

那么

$$D = L'_{\text{Hinge}}(\mathbf{x}', \mathbf{y}', \mathbf{w}, \mathbf{b}) \geq 0.$$

2) 如果

$$m + \sum_{i \in S} (\bar{y}_i' - y_i')(\mathbf{w}_i^T \phi_i(\mathbf{x}') + b_i) + \sum_{i \in S} \sum_{j \in N_i} (\bar{y}_i' \bar{y}_j' - y_i' y_j') \mathbf{w}_{ij}^T \phi_j(\mathbf{x}') > 0,$$

那么

$$\begin{aligned} D &= -m + \sum_{i \in S} \left[\max(0, m_i - 2y_i'(\mathbf{w}_i^T \phi_i(\mathbf{x}') + b_i)) - (\bar{y}_i' - y_i')(\mathbf{w}_i^T \phi_i(\mathbf{x}') + b_i) \right] \\ &\quad + \sum_{i \in S} \sum_{j \in N_i} \left[\max(0, m_{ij} - 2y_i' y_j' \mathbf{w}_{ij}^T \phi_j(\mathbf{x}')) - (\bar{y}_i' \bar{y}_j' - y_i' y_j') \mathbf{w}_{ij}^T \phi_j(\mathbf{x}') \right]. \end{aligned}$$

令

$$D_i = \max(0, m_i - 2y_i'(\mathbf{w}_i^T \phi_i(\mathbf{x}') + b_i)) - (\bar{y}_i' - y_i')(\mathbf{w}_i^T \phi_i(\mathbf{x}') + b_i)$$

$$D_{ij} = \max(0, m_{ij} - 2y_i' y_j' \mathbf{w}_{ij}^T \phi_j(\mathbf{x}')) - (\bar{y}_i' \bar{y}_j' - y_i' y_j') \mathbf{w}_{ij}^T \phi_j(\mathbf{x}'),$$

那么

$$D = -m + \sum_{i \in S} D_i + \sum_{i \in S} \sum_{j \in N_i} D_{ij}.$$

分两种情况讨论 D_i ,

2.1) 如果

$$m_i - 2y_i^t(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i) \leq 0,$$

那么

$$\begin{aligned} D_i &= -(\bar{y}_i^t - y_i^t)(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i) \\ &= \begin{cases} 0 & \text{if } \bar{y}_i^t = y_i^t \\ 2y_i^t(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i) \geq m_i & \text{if } \bar{y}_i^t = -y_i^t \end{cases}. \end{aligned}$$

2.2) 如果

$$m_i - 2y_i^t(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i) > 0,$$

那么

$$\begin{aligned} D_i &= m_i - (\bar{y}_i^t + y_i^t)(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i) \\ &= \begin{cases} m_i - 2y_i^t(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i) > 0 & \text{if } \bar{y}_i^t = y_i^t \\ m_i & \text{if } \bar{y}_i^t = -y_i^t \end{cases}. \end{aligned}$$

由 2.1)和 2.2)可知

$$D_i \begin{cases} \geq 0 & \text{if } \bar{y}_i^t = y_i^t \\ \geq m_i & \text{if } \bar{y}_i^t = -y_i^t \end{cases}.$$

同理可得

$$D_{ij} \begin{cases} \geq 0 & \text{if } \bar{y}_i^t \bar{y}_j^t = y_i^t y_j^t \\ \geq m_{ij} & \text{if } \bar{y}_i^t \bar{y}_j^t = -y_i^t y_j^t \end{cases}.$$

由以上两个式子和已知条件 $m \leq \sum_{i \in S, \bar{y}_i^t = -y_i^t} m_i + \sum_{i \in S, j \in N_i, \bar{y}_i^t \bar{y}_j^t = -y_i^t y_j^t} m_{ij}$ 可得

$$\begin{aligned} D &= -m + \sum_{i \in S, \bar{y}_i^t = y_i^t} D_i + \sum_{i \in S, \bar{y}_i^t = -y_i^t} D_i + \sum_{i \in S, j \in N_i, \bar{y}_i^t \bar{y}_j^t = y_i^t y_j^t} D_{ij} + \sum_{i \in S, j \in N_i, \bar{y}_i^t \bar{y}_j^t = -y_i^t y_j^t} D_{ij} \\ &\geq -m + \sum_{i \in S, \bar{y}_i^t = -y_i^t} D_i + \sum_{i \in S, j \in N_i, \bar{y}_i^t \bar{y}_j^t = -y_i^t y_j^t} D_{ij} \\ &\geq 0 \end{aligned}.$$

因此, 在 1)和 2)两种情况下, 均有 $D \geq 0$, 原命题得证。■

参考文献

- [1] J. Jeon, V. Lavrenko and R. Manmatha. Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In Proceedings of the 26th annual international ACM SIGIR conference, pages 119-126, 2003.
- [2] V. Lavrenko, R. Manmatha, and J. Jeon. A Model for Learning the Semantics of Pictures. In Proceedings of conference on Neural Information Processing Systems, 2003.
- [3] S. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. In Proceeding of IEEE conference on Computer Vision and Pattern Recognition, pages 1002-1009, 2004.
- [4] M. Szummer and R. Picard. Indoor-Outdoor Image Classification. In Proceedings of IEEE international workshop on Content-Based Access of Image and Video Database, pages 42-51, 1998.
- [5] A. Vailaya, A. Jain, and H. Zhang. On Image Classification: City vs. Landscape. Pattern Recognition, 31:1921-1936, Dec. 1998.
- [6] C. Cusano, C. Ciocca and R. Schettini. Image Annotation using SVM. In Proceedings of Internet Imaging V, 5304:330-338, 2003.
- [7] G. Carneiro, A. Chan, P. Moreno and N. Vasconcelos. Supervised Learning of Semantic Classes for Image Annotation and Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(3):394-410, 2007.
- [8] P. Duygulu, K. Barnard, J. Freitas and D. Forsyth. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In Proceedings of European Conference on Computer Vision, 2002.
- [9] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu and S. Ma. Dual Cross-Media Relevance Model for Image Annotation. In Proceedings of the 15th International Conference on Multimedia, 605-614, 2007.
- [10] X. Zhou, M. Wang, Q. Zhang, J. Zhang and B. Shi. Automatic Image Annotation by an Iterative Approach: Incorporating Keyword Correlations and Region Matching. In Proceedings of the 6th International Conference on Image and Video Retrieval, 25-32, 2007.
- [11] N. Rasiwasia and N. Vasoncelos. Holistic Context Modeling using Semantic Co-occurrences. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1889-1895, 2009.

- [12] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei and H. Zhang. Correlative Multi-Label Video Annotation. In Proceedings of the 15th International Conference on Multimedia, 17-26, 2007.
- [13] I. Tsochantaridis, T. Hofmann, T. Joachims and Y. Altun. Support Vector Machine Learning for Interdependent and Structured Output Spaces. In Proceedings of ACM International Conference on Machine Learning, 2004.
- [14] R. Kindermann and J. L. Snell. Markov Random Fields and Their Applications. American Mathematical Society, 1980.
- [15] Y. Xiang, X. Zhou, T.S. Chua and C.W. Ngo. A Revisit of Generative Model for Automatic Image Annotation using Markov Random Fields. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1153-1160, 2009.
- [16] Y. Xiang, X. Zhou, Z. Liu, T.S. Chua and C.W. Ngo. Semantic Context Modeling with Maximal Margin Conditional Random Fields for Automatic Image Annotation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010.
- [17] S.Z. Li. Markov Random Field Modeling in Computer Vision. Springer-Verlag Press, 1995.
- [18] J. Lafferty, A. McCallum and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of ACM International Conference on Machine Learning, 2004.
- [19] C. Wang, L. Zhang and H. Zhang. Scalable Markov Model-based Image Annotation. In Proceedings of the International Conference on Content-based Image and Video Retrieval, 113-118, 2008.
- [20] C. Yang, M. Dong and J. Hua. Region-based Image Annotation using Asymmetrical Support Vector Machine-based Multiple-Instance Learning. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2057-2063, 2006.
- [21] A. Makadia, V. Pavlovic and S. Kumar. A New Baseline for Image Annotation. In Proceedings of the European Conference on Computer Vision, 316-329, 2008.
- [22] M. Guillaumin, T. Mensink, J. Verbeek and C. Schmid. TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-Annotation. In Proceedings of the International Conference on Computer Vision, 2009.
- [23] A. Oliva and A. Torralba. The Role of Context in Object Recognition. Trends in Computer Science, 11(12):520-527, 2007.
- [24] L. Wolf and S. Bileschi. A Critical View of Context. International Journal of Computer Vision, 69(2):251-261, 2006.

-
- [25] J. Poyway, K. Wang, B. Yao and S. Zhu. A Hierarchical and Contextual Model for Aerial Image Understanding, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1-8, 2008.
- [26] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora and S. Belongie. Objects in Context. In Proceedings of IEEE International Conference on Computer Vision, 1-8, 2007.
- [27] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145-175, 2001.
- [28] F.F. Li and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 542-531, 2005.
- [29] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721-741, 1984.
- [30] H. Derin and H. Elliott. Modeling and Segmentation of noisy and textured Image using Gibbs Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):39-55, 1987.
- [31] B. Micusik and T. Pajdla. Multi-label Image Segmentation via Max-sum Solver. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1-6, 2007.
- [32] Y. Li, Y. Tsing, Y. Genc and T. Kanade. Object Detection using 2D Spatial Ordering Constraints. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005.
- [33] X. He, R. Zemel and M. Peripin. Multiscale Conditional Random Fields for Image Labeling. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2:695-702, 2004.
- [34] A. Quattoni, M. Collins and T. Darrell. Conditional Random Fields for Object Recognition. In Proceedings of conference on Neural Information Processing Systems, 2004.
- [35] L. Cao, J. Luo, H. Kautz and T. Huang. Annotating Collections of Photos using Hierarchical Event and Scene Models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1-8, 2008.
- [36] S. Kumar and M. Hebert. Discriminative Fields for Modeling Spatial Dependencies in Natural Images. In Proceedings of Conference on Neural Information Processing Systems, 2004.
- [37] B. Taskar, C. Guestrin and D. Koller. Max-Margin Markov Networks. In Proceedings of conference on Neural Information Processing Systems, 2003.

- [38] J. Besag. On the Statistical Analysis of Dirty Pictures. Journal of the Royal Statistical Society, 1986.
- [39] R. Jin, Y. Chai and L. Si. Effective Automatic Image Annotation via a Coherent Language Model and Active Learning. In Proceedings of the 12th International Conference on Multimedia, 892-899, 2004.
- [40] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato and F. Huang. A Tutorial on Energy-based Learning. Predicting Structured Data, MIT Press, 2006.
- [41] V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.
- [42] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60(2):91-110, 2004.
- [43] Y. Jiang, C.W. Ngo and J. Yang. Towards Optimal Bag-of-features for Object Categorization and Semantic Video Retrieval. In Proceedings of ACM International Conference on Image and Video Retrieval, 494-501, 2007.
- [44] Y. Wang and S. Gong. Refining Image Annotation using Contextual Relations Between Words. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, 425-432, 2007.
- [45] M. Srikanth, J. Varner, M. Bowden and D. Moldovan. Exploiting Ontologies for Automatic Image Annotation. In Proceedings of the 28th Annual International ACM SIGIR Conference, 552-558, 2005.
- [46] Y. Jin, L. Khan, L. Wang and M. Awad. Image Annotation by Combining Multiple Evidence & WordNet. In Proceedings of the 13th Annual ACM International Conference on Multimedia, 706-715, 2005.
- [47] M. Wang, X. Zhou and T.S. Chua. Automatic Image Annotation via Local Multi-Label Classification. In Proceedings of the International Conference on Content-based Image and Video Retrieval, 17-26, 2008.
- [48] Y. Wang, T. Mei, S. Gong and X. Shen. Combining Global, Regional and Contextual Features for Automatic Image Annotation. Pattern Recognition, 42(2):259-266, 2009.
- [49] N. Haering, Z. Myles and N. Lobo. Locating Dedicuous Trees. In Proceedings of Workshop in Content-based Access to Image and Video Libraries, 18-25, 1997.
- [50] D. Forsyth and M. Fleck. Body Plans. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 678-683, 1997.
- [51] Y. Li and L. Shapiro. Consistent Line Clusters for Building Recognition in CBIR. In Proceedings of International Conference on Pattern Recognition, 3:952-956, 2002.
- [52] J. Fan, Y. Gao and H. Luo. Hierarchical Classification for Automatic Image Annotation.

- In Proceedings of the 30th annual International SIGIR Conference, 111-118, 2007.
- [53] M. Fink and P. Perona. Mutual Boosting for Contextual Inference. In Proceedings of Conference on Neural Information Processing Systems, 2004.
- [54] S. Lazebnik, C. Schmid and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2169-2178, 2006.
- [55] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. M. Blei and M. I. Jordan. Matching Words and Pictures. Journal of Machine Learning Research, 3:1107-1135, 2003.
- [56] F. Monay and D. Gatica-Perez. PLSA-based Image Auto-Annotation: Constraining the Latent Space. In Proceedings of the 12th Annual ACM International Conference on Multimedia, 348-351, 2004.
- [57] O. Yakhnenko and V. Honavar. Annotating Images and Image Objects using a Hierarchical Dirichlet Process Model. In Proceedings of the 9th International Workshop on Multimedia Data Mining, 1-7, 2008.
- [58] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li and D. N. Metaxas. Automatic Image Annotation using Group Sparsity. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010.
- [59] D. Putthividhya, H. T. Attias and S. S. Nagarajan. Topic Regression Multi-Modal Latent Dirichlet Allocation for Image Annotation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010.
- [60] O. Maron and T. Lozano-Perez. A Framework for Multiple-Instance Learning. In Proceedings of Conference on Neural Information Processing Systems, 570-576, 1998.

攻读学位期间作者的研究成果

1. 参与科研项目

国家自然科学基金项目：基于超平面查询的 Web 图像数据库索引及主动学习研究
国家自然科学基金项目：基于语言模型的图像数据库自动标注与多模式检索研究

2. 已发表和录用论文

- [1] **Yu Xiang**, Xiangdong Zhou, Zuotao Liu, Tat-Seng Chua and Chong-Wah Ngo. Semantic Context Modeling with Maximal Margin Conditional Random Fields for Automatic Image Annotation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010. Accepted.
- [2] **Yu Xiang**, Xiangdong Zhou, Tat-Seng Chua and Chong-Wah Ngo. A Revisit of Generative Model for Automatic Image Annotation using Markov Random Fields. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1153-1160, 2009.
- [3] Hongtao Xu, Xiangdong Zhou, Mei Wang, **Yu Xiang**, Baile Shi. Exploring Flickr's Related Tags for Semantic Annotation of Web Images. In Proceedings of the ACM International Conference on Image and Video Retrieval, 1-8, 2009.
- [4] Hongtao Xu, Xiangdong Zhou, Lan Lin, **Yu Xiang**, Baile Shi. Automatic Web Image Annotation via Web-Scale Image Semantic Space Learning. In Proceedings of the Joint International Conferences on Asia-Pacific Web Conference (APWeb) and Web-Age Information Management (WAIM), 211-222, 2009.
- [5] 郝国煜, 向宇, 周向东, 施伯乐. 支持向量机 Top-k 查询的特征空间近邻索引. 第二十五届中国数据库学术会议(NDBC), 2008.
- [6] 许红涛, 周向东, 向宇, 施伯乐. 一种自适应的 Web 图像语义自动标注方法. 软件学报, 已录用.

致谢

我衷心地感谢我的导师，周向东副教授。周老师勤奋严谨的治学态度，一丝不苟的科研精神深深地影响着我。感谢周老师一直以来的关心和帮助，您的言传身教坚定了我继续深造，探索科学知识的决心。感谢周老师教会了我做学问的态度和方法，您的悉心指导给予我很大的帮助和启发。

感谢施伯乐教授对我循循善诱的教诲和帮助。施老师在学习和生活上给予了我无微不至的关心和爱护，我将永远铭记于心。

感谢新加坡国立大学的 Chua Tat-Seng 教授和香港城市大学的 Ngo Chong-Wah 教授对于我的论文所提出的建设性意见。两位教授在英文论文写作上给予了我的许多帮助。

作为周老师研究小组中的一员，我与其他小组成员互相讨论想法，分享经验。感谢王梅，许红涛，袁进，郝国煜，刘作涛，产文，李真超，傅德基和纪传俊给予我的帮助，与你们的讨论使我获益匪浅。

感谢父母在生活上的关爱和精神上的鼓励，使在外求学的我感受到家的温暖。感谢我的家人和朋友的支持，鼓励和陪伴。

论文独创性声明

本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。论文中除了特别加以标注和致谢的地方外，不包含其他人或其它机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中作了明确的声明并表示了谢意。

作者签名：_____ 日期：_____

论文使用授权声明

本人完全了解复旦大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。保密的论文在解密后遵守此规定。

作者签名：_____ 导师签名：_____ 日期：_____