中国科学技术大学

学士学位论文

多模态场景下的自监督学习算法研究

作者姓名: 付哲仁

学科专业: 电子信息工程

导师姓名: 毛震东 特任研究员

完成时间: 二〇二〇年五月二十八日

University of Science and Technology of China

A dissertation for bachelor's degree

Self-supervised representation learning algorithm on multimodal scene

Author: Fu Zheren

Speciality: Electronic Information Engineering

Supervisor: Prof. Zhendong Mao

Finished time: May 28, 2020

致　谢

在中国科学技术大学的本科四年生活与学习中，受到了老师和同学们的许多帮助，没有他们就没有我现在的成就，在完成论文之际，请容许我对他们表达真诚的感谢。

首先感谢导师毛震东老师大四学期以来的监督与帮助，为我提供了良好的学习环境与研究资源，对我进行了细心的指导，将研究小白的我带入计算机视觉研究领域，为我将来的研究生学习打下坚实基础。感谢班主任方老师四年来对班级的组织管理，您辛苦了，信息学院 1604 班在您的带领下，成为一个和谐温馨的大家庭。

感谢实验室的师兄师姐们，当我遇到研究困难时给予我科研上的关心，这样我才能解决一个又一个难题，未来我们一起努力。感谢室友们生活上的照顾，四年来的相处很愉快。感谢中国科大，我以自己是科大学生而自豪。

最后，感谢我家人一贯的鼓励和支持，你们是我追求学业的坚强后盾。

目 录

中文内容摘要

深度学习在越来越多的计算机视觉任务上取得很好的效果，背后通常依赖大量人工标注信息，获取需要很高的成本。而自监督学习算法只需要利用未标注的数据，根据数据之间的相关性自动地抽取出合适的“监督信息”，指导模型进行有效的学习，并能运用到各种下游任务中去。针对视觉领域的自监督算法，利用图像之间的相关性，构造各种各样的辅助任务训练模型，但图像是单模态的数据，且自身的表示信息弱，语义层级低，因此如何挖掘更深层的监督信号，是自监督学习研究面对的主要挑战。

由于文本自身的语义性要强于图像，多模态数据的信息也存在互补性，因此本文选择在多模态场景下构造自监督学习算法，利用多模态数据之间的相关性为算法提供监督信息，打破单模态信息载体的局限性。网络上的多媒体数据通常是以图片-文本形式成对出现的，二者语义内容接近，并且文本的语义特征可以单独训练获取。因此我们以文本的语义信息作为监督信号，把预先计算好文本表示特征看作对应图片的语义标签，设计合适的损失函数，训练视觉特征表示模型。

本文使用了从维基百科网页，社交媒体平台上收集的免费公开的数据集，作为无标注训练数据，结合不同的语言模型和度量学习损失函数，训练深度卷积神经网络。最后为了验证算法的可靠性，做了大量对比实验。发现本文提出多模态场景下的自监督学习算法框架，能超过大部分单模态的自监督算法，帮助模型学习到更有分辨力的视觉特征表示，并在下游的图片分类，目标检测，语义分割，甚至跨模态检索等任务上具有更强的泛化能力，我们有理由相信未来基于多模态数据的自监督学习算法将成为主流研究方向。

关键词：人工智能；机器学习；自监督学习；表示学习；多模态内容理解

Abstract

Deep learning has achieved good results in more and more computer vision tasks. It usually relies on a large number of manual annotation information, which requires a high cost to obtain. The self-supervised learning algorithm only needs to use the unlabeled data, according to the correlation between the data, automatically extract the appropriate ”supervised information”, guide the model for effective learning. And it can be used in various downstream tasks. In view of the self-supervised algorithm of computer vision, a variety of auxiliary task training models are constructed by using the correlation between images. But the image is single-modal data so that its representation information is weak and its semantic level is low. How to mine deeper supervised signals is the main challenge of self-supervised learning research.

Because the semantics of text is stronger than that of image, and the information of multimodal data is complementary. This paper chooses to construct a self-supervised learning algorithm in multimodal scene, using the correlation between multimodal data to provide supervision information for the algorithm, breaking the limitations of single-mode information carrier. Multimedia data on the network usually appears in pairs in the form of picture text. The semantic content of the two is similar, and the semantic features of the text can be acquired by training alone. Therefore, we take the semantic information of text as the monitoring signal, regard the pre calculated text representation features as the semantic labels of the corresponding pictures, design the appropriate loss function, and train the visual feature representation model.

In this paper, we use the free and open data set collected from Wikipedia web page and social media platform as training data without annotation. we train the deep convolution neural network with different language models and learning loss measurement functions. Finally, in order to verify the reliability of the algorithm, a lot of comparative experiments are done. It is found that the self-supervised learning algorithm framework proposed in this paper can surpass most of the single-mode self-supervised algorithms, help the model learn more discriminative visual feature representation which have stronger generalization ability in downstream task like image classification, target detection, semantic segmentation, and even cross-modal retrieval. We have reason

to believe that the self-supervised learning algorithm based on multimodal data will become the main research direction in the future.

Key Words: artificial intelligence; machine learning; self-supervised learning; representation learning; multimodal context understanding

第一章　绪论

第一节　研究背景与意义

图像分类、目标检测、语义分割等计算机视觉相关问题，目前最好的解决的方法都是基于有监督的机器学习或深度学习算法，在标注数据的支撑下训练深度神经网络模型。然而这些从标签数据中学习的方法背后都极度依赖于庞大的人工标注数据，数据通常容易获得，但针对不同的任务给它们打上特定的标签是一项耗时费力的工作。对于大部分公司和研究机构，由于经济成本太高，短时间内很难获取到某个研究方向的大量有标注数据，在医学这种对技术要求较高的领域，人工标注的过程也很容易出错。

若能从未标记的数据中学习到有帮助的信息，可以显著降低将机器学习算法部署到新应用的成本，从而增强它们在现实世界中的影响力。自监督学习就是一套不需要标签也能从数据中学习的算法框架，它的思想是从原始的数据中寻找相关性约束，构建对应的学习任务，指导模型学习语义表示信息。它的出现能在一定程度上缓解深度学习算法严重依赖数据标签的问题，逐渐成为机器学习领域的重要研究方向。

第二节　自监督学习的进展与挑战

在计算机视觉领域，自监督学习的研究主要集中在根据图像的自身的相对关系，设计各种合适的学习任务：如预测两个图像块之间的位置关系 [9]，图像着色 [75]，图像拼图 [46]，预测图像的旋转角度 [13]，图像特征聚类等 [3]。自监督算法使模型学习到对视觉任务有帮助的表示特征，并可以应用到许多其他的机器学习领域中，如少样本学习 [59]、半监督学习 [74]、训练生成对抗网络 [6]、多任务学习 [10] 等方向。纵观自监督学习近今年来的发展，发现监督信息的挖掘越来越深入，学习任务的设计越来越复杂，但这些方法有共同的特点：1. 视觉特征需要被卷积神经网络提取，并去解决各种设计好的学习任务。2. 学习任务的监督信息或伪标签可以根据图像的属性与关系获取。

目前自监督学习绝大部分工作都是基于公开的图像分类数据集 ImageNet[7] 上做的，仅限于视觉模态的信息被利用，其他模态的信息 (语音，文本) 都未被加入。随着算法性能的不断提高，如何进一步挖掘图像内部或图像之间的关联性，

寻找更深层次的潜在约束作为指导模型训练的监督信息的工作也越来越难进行。这些都是当前自监督学习研究面临的挑战。

第三节　基于多模态场景的自监督学习算法

我们发现，人类产生的图像数据标签与文本有紧密关系，即标签由不同粒度的文本信息形式的语义实体组成，如一个单词定义单个物体的分类标签，多个单词或短语对应场景的描述标签。同时与原始图像像素相比，单词无疑代表了更多的高级语义概念，因此我们考虑把文本模态的信息加入到自监督算法中。ImageNet 数据集本身并不包含文本信息，但图片与文本成对出现的数据在网络空间中随处可见，大规模地存在于新闻网站，百科全书条目，社交分享应用等多媒体平台中。数据的视觉和文本内容相辅相成，为多模态内容理解提供更深层的语义信息。我们希望能利用这些免费获取、包含噪声的无标注多媒体数据来构建自监督算法，训练视觉模型。

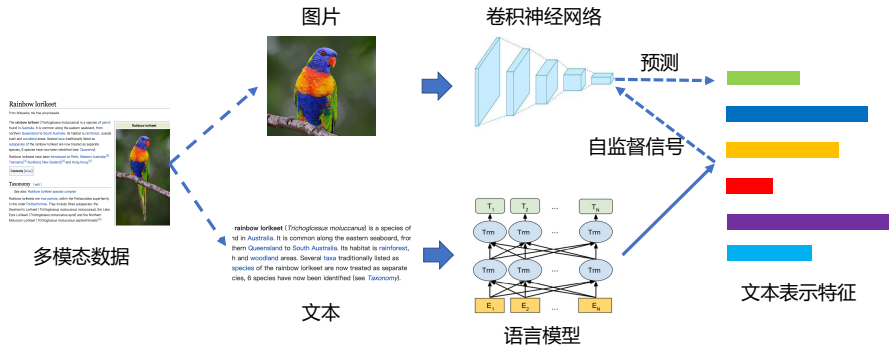

图 1.1　算法整体框架，用图片相关的文本语义信息监督卷积神经网络进行学习

基于图片-文本对的自然共生关系，我们假设二者之间存在一定深度的语义相关性，把文本中的语义内容作为学习视觉特征的监督信号，设计对应的自监督表示学习任务，指导网络模型学习视觉特征。算法框架整体如图 1.1所示，具体来说，首先选择合适的文本编码器，用预训练的方式提取文本特征，再训练视觉编码器 (深度卷积神经网络)，把图像投影到上一步的构建好的文本语义空间，学习视觉特征。我们希望由卷积神经网络提取的视觉特征要与对应文本特征尽量对齐，这样神经网络模型也具有一定的语义表示能力。

第四节　实验结果及本文主要贡献

我们从维基百科与社交媒体平台获取大量无标注数据来训练网络模型，并尝试了多种语言模型作为文本编码器，同时结合了不同类型的损失函数，做了大量对比实验。为了验证本文的自监督算法有效性，在图像分类任务上训练线性分类器，把预训练模型迁移到目标检测、语义分割等下游任务中，实验表明，我们的多模态场景下的自监督算法能性能与主流的方法不相上下，在特定任务上甚至更好，并且我们的算法框架能应用到多模态检索任务中，且效果能接近主流的有监督学习方法。本文的主要工作和贡献可总结为如下几点：

- 提出了一种在多模态场景下利用图像与语义文本之间的相关性，完成对视觉特征的自监督表示学习方法。
- 尝试了不同种类的多模态无标注数据集、文本编码器、度量损失函数，寻找最适合的该算法框架的组合，综合比较了不同组合间学习特征的差异。
- 在多种下游任务与数据集上进行实验，证明基于文本内容学习训练的视觉特征，具有语义特性，在图像分类、目标检测等基准任务中，具有与最近的自监督和无监督算法相当的性能，并且可直接应用在跨模态检索领域中。

第二章　相关工作的研究综述

第一节　自监督学习

自监督学习是无监督学习领域的重要方向，顾名思义，就是从未标注的大量数据自动地抽取出“监督信息”，从而指导算法学习有用的特征，属于无监督学习的一部分。自监督学习的网络可以用于特定下游任务的网络初始化，或直接将网络的输出作为下游任务的特征输入。目前自监督学习的关键研究点在于：1. 如何从未标注的数据挖掘有效的自监督信息，即从未标注的数据中挖掘什么样的信息来指导网络的训练; 2. 针对特定的下游任务，设计有效的学习任务 (pretext task)，即不同下游任务依赖特征的语义层级有所不同，需要设计相应的学习任务来驱动网络学习提取合适的特征。

近几年该在计算机视觉领域，自监督学习备受关注，涌现了很多有创造性、有价值的工作。监督信息的挖掘越来越深入，学习任务设计越来越复杂，如预测图像的颜色 [75, 37]，预测图像旋转角度 [13]，预测图像子块相对位置 [9, 46]，视频帧的时序预测 [47, 45]，数据集特征聚类 [3, 4]，图像近邻关系 [56]，最近的基于正负样本间对比学习 [29, 61, 70] 的方法十分成功，取得了当前最好的效果，是目前的主流研究方向。

在使用视觉信息的同时，越来越多工作尝试引入其他模态信息，如文本 [43]，声音 [48] 等。也有将自监督学习和半监督学习 [74]，小样本学习 [59]，多任务学习 [10] 等其他方向研究相结合，进一步提升效果的工作。于此同时，也有越来越多与实际应用相结合的工作出现，如多模态预训练模型 [60] 等，甚至有些自监督学习算法 [25, 5] 的效果在某些下游任务上可超过有监督学习。

我们的工作是在多模态场景下的，近几年也有很多相似的工作 [17, 15, 50, 52]，也有一些结合弱监督学习的方法，[71] 结合 10 亿的无标注图片和 ImageNet[32] 标注图片，提出基于教师网络的半监督学习框架，获得比全监督更好的效果，[42] 探究无标注可用数据的极限，选取社交媒体上图片的话题标签 (hashtag) 作为标注信息预训练模型，在下游分类任务上获得目前最高的准确率，[72] 提出用自监督学习预训练好的网络帮助海量噪声生成伪标签的方法，增强无标注数据的利用效率。这些方法充分挖掘互联网上巨量、繁杂的多模态数据，以多模态内容语义不变性作为目标约束，引入视觉相关的文本语义特征作为监督信息，即先学习文本特征并固定，再来学习基于多模态语义对齐的文本-图像

联合表示，优化视觉网络模型，从而指导其学习到有效的数据表示。

第二节　跨模态表示学习

跨模态检索，图像描述，视觉问答等任务要求同时利用相关图像与自然语言，一个好的解决方法是在一个特征空间里建立起视觉和文本的联合特征，多模态的信息直接在统一的空间里操作比较。跨模态表示学习研究如何将不同模态的特征映射到一个统一的表示空间，这些联合特征可直接被用于以图搜文字、以文字搜图等应用中。

跨模态表示学习早期一类方法是基于典型相关分析 [24](CCA) 及其扩展的工作 [58, 19]，是利用综合变量之间的相对关系来反映两对变量之间的整体相关性的多元统计分析方法，在数据挖掘领域也有应用。区别于 CCA，早期的另一类方法是基于排序损失函数的联合特征学习的工作，比如 WSABIE[69] 和 DeViSE[12]，直接用单个排序损失函数，在文本与图像特征之间学习一个线性变换模型。也有文章提出利用附加约束 [67] 的双向排序损失函数来优化模型 [32, 33]。还有基于二值模型的多模态的哈希算法，研究如何有效的把不同模态的数据映射到一个二进制汉明空间，如将语义标签集成到哈希学习过程中的 SCM[77]，与 CCA 相结合，显式地结合了视觉特征和文本的依赖性的 CCA-3V[18]。

之后有人提出一个新的跨模态匹配问题的正则化框架 LCFS[66](Learning Coupled Feature Spaces)，它将耦合线性回归、l2 范数和跟踪范数统一为一个通用的最小化公式，从而可以同时进行子空间学习和耦合特征选择。此外，在 [65] 中他们将这个框架扩展到两种以上的情况，获得一个扩展版本为 JFSSL(Joint Feature Selection and Subspace Learning)，还有把对抗学习 [64] 用在跨模态检索上。

近几年与深度学习相结合的方法在此任务上也有广泛使用，比如利用神经网络模型作为特征提取单元，图像使用在 ImageNet[32] 数据集上预训练的卷积神经网络 (AlexNet[35]，ResNet[26]), 文本使用循环神经网络 (LSTM[27])。一个相似的想法是利用图片标题为语义检索学习全局的视觉表示，在 [21] 中使用基于 BOW 的 tf-idf[31] 来表示图片标题，以此训练一个卷积神经网络 (CNN) 网络，通过最小化三元组损失函数 [28] 作为图片的语义相似性测量。[67] 提出了一种通过训练神经网络在同一空间嵌入 Word2Vec[44] 文本表示和 CNN 提取特征，来学习图像与文本联合嵌入的方法，用于图文之间的相互检索。

除了语义检索之外，图像文本联合表示也被应用于更为具体的应用中。[51]

中使用 LDA[1] 学习图像-文本联合嵌入，并仅使用视觉信息生成图像的语境化词汇。如 [20] 基于 WordNet[①]提供的图形分类法，在语义空间中嵌入单词相关的图像信息，完成文本识别的任务。特别的，[57] 提出了一种基于食品图像及其配方文字的联合嵌入以识别成分的方法，使用 Word2Vec 和 LSTM 网络对成分名称和烹饪说明文本进行编码，并使用 CNN 从相关图像中提取视觉特征。

然而在跨模态任务获得大量的图片-文本样本对需要耗费大量人力去清洗、标注数据，成本很高。我们的自监督学习算法只需要无标签的数据，不同于以前的图像文本嵌入方法，省去了数据标注的麻烦，最后模型也学习到了具有一定的泛化性和辨识性的特征，甚至在一些任务上能与有监督学习算法相媲美。

[①]https://wordnet.princeton.edu/

第三章　多模态的自监督算法模型

本文提出的基于图文匹配的多模态自监督算法，如图 3.1所示，无标注多模态数据集存在的内容语义相关的图片-文本对，分别用视觉编码器与文本编码器提取特征，得到对应模态数据的表示，我们希望表示特征映射到同一个嵌入空间，由于文本编码器可以单独训练，故以文本特征的语义信息作为对应图片的监督信号，指导视觉编码器 (卷积神经网络模型) 训练，学习可靠的视觉特征。

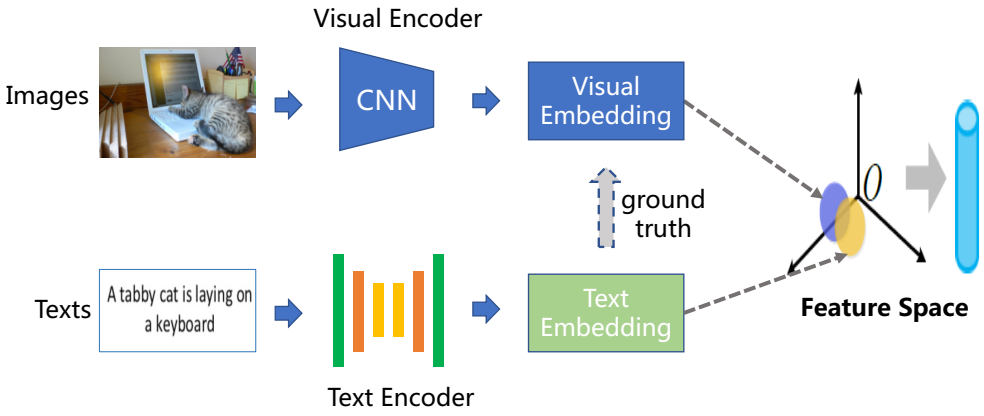

图 3.1　算法框架，利用文本编码器提取特征，作为视觉编码器训练的监督信号，希望二者输出的特征空间尽可能靠近

第一节　算法框架详细介绍

对于给定的多模态数据集 C，相关的图片文本对为 $\{x_i, t_i\} \in C,\ i = 1, \cdots N$，$x_i$ 表示第 i 个图片样本，t_i 表示第 i 个文本样本，$N = |C|$，表示数据集的大小。我们利用文本特征编码器 $\varphi(t_i) \in \mathbb{R}^{d_t}$ 来提取语言表示信息，并用图像特征编码器 $\phi(x_i) \in \mathbb{R}^{x_t}$ 来来提取视觉表示信息，d_t, d_x 是分别表示文本特征向量与图片特征向量的维度，本文中二者相等，记为 d。我们希望利用文本的语义信息作为监督信号，因此把文本特征作为图像的标签，$y_i = \varphi(t_i)$，用它来指导视觉特征编码器 $\phi(\cdot)$ 的优化，使编码器输出的特征 $f_i = \phi(x_i)$ 有一定分辨性。

基于相关图片和文本对应该具有语义相似性的假设原则，我们希望经过各自编码后的特征也具有相似性，特征之间的相关性应该尽可能大。定义距离度量

函数 $\mathbb{D}$ 与相似度量函数 $\mathbb{S}$，我们的优化目标是希望：

$$argmin\frac{1}{N}\sum_{i=1}^{N}\mathbb{D}(f_i, y_i) \quad or \quad argmax\frac{1}{N}\sum_{i=1}^{N}\mathbb{S}(f_i, y_i) \tag{3.1}$$

$\mathbb{S}$ 在度量学习中，一般选用余弦相似函数，来测量两个特征向量相似性，余弦相似又称为余弦距离，但余弦距离大小与特征相似性呈反比，因此实际上属于相似度量函数，因此优化目标可写成：

$$argmax\ \frac{1}{N}\sum_{i=1}^{N}\frac{\langle\phi(x_i), \varphi(t_i)\rangle}{\|\phi(x_i)\| \cdot \|\varphi(t_i)\|} \tag{3.2}$$

很多度量学习的损失函数可帮助优化各种的表示学习问题，增加正样本对之间对相似性，同时减少负样本对之间的相似性，使类内特征聚集，类间特征分散。负样本可以用各种困难样本挖掘策略，寻找难以优化的样本，且负样本对之间的不相似程度常常存在一个间隔 m(超参数) 来控制。本文中我们选用对比损失 (contrastive loss[23]) 和三元损失 (triplet loss[28]) 来优化，在包含 M 个图片-文本对的小批次数据 (mini-batch) 中循环构造正负样本对，给定视觉特征 $\phi(x_i)$，对应的文本特征 $\varphi(t_i)$ 为正样本，其余的全为负样本，对比损失有：

$$Loss = \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1,\ j\neq i}^{M}[(1 - S_{ii}) + [m - S_{ij}]_+] \tag{3.3}$$

三元损失有：

$$Loss = \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1,\ j\neq i}^{M}[m + S_{ij} - S_{ii}]_+ \tag{3.4}$$

其中 $[x]_+$ 为 hinge 折页函数，又可以写成 $max(x, 0)$，S_{ij} 表示计算 x_j 与 t_j 对应特征的余弦相似度，即：

$$S_{ij} = cosine\_similarity(\phi(x_i), \varphi(t_j)) \tag{3.5}$$

$\mathbb{D}$ 一般选取的 L2 归一化的欧式距离函数，等价于余弦距离，计算相对复杂，实际中基本用余弦距离来计算。特别的，对带有概率含义的特征可采取 KL 散度函数来测量两个概率分布之间的距离，又称为交叉熵损失函数，因此优化目标可写成：

$$argmin\ \frac{1}{Nd}\sum_{i=1}^{N}[-\sum_{j=1}^{d}\varphi(t_i)_j \cdot \log\phi(x_i)_j] \tag{3.6}$$

对于普通的特征向量，赋予向量每个位置元素概率含义，即添加元素大小 0-1 的约束后，也可以用二元交叉熵损失来优化。记 $\sigma(\cdot)$ 为元素操作的 sigmoid

函数，$p_{ij} = \sigma(\varphi(t_i)_j)$，$\hat{p}_{ij} = \sigma(\phi(x_i)_j)$，其中 $p_{ij}, \hat{p}_{ij} \in \mathbb{R}^d$，$d$ 是特征向量的维度大小，因此优化目标（损失函数）可写成：

$$Loss = -\frac{1}{Nd}\sum_{i=1}^{N}\sum_{j=1}^{d}\left[p_{ij}\log\hat{p}_{ij} + \left(1-p_{ij}\right)\log\left(1-\hat{p}_{ij}\right)\right] \tag{3.7}$$

本文的视觉编码器 $\phi(\cdot)$ 选用为分类任务设计的深度卷积神经网络 (CNN) 模型，是我们需要训练与优化的对象。文本编码器 $\varphi(\cdot)$ 选择典型的语言模型。语言模型独自预训练好再使用，并在上述的自监督算法中冻结模型的参数，使输出的文本特征固定，当然语言特征也可与视觉特征一样动态更新，但这样训练过程极为不稳定，模型难以收敛。因此语言模型做预训练处理，文本特征只是作为监督信息参与 CNN 的训练但不被优化更新。

第二节　文本编码器

因为文本特征是监督信息，指导视觉编码器训练，因此选择一个合适的语言模型来编码文本的语义特征十分关键。根据模型结构与处理文本的方式，文本矢量化方法多种多样，有些方法是面向单个单词的矢量化，而另一些是面向全文或段落的矢量化。为了获得性能最好的文本特征表示，我们选用不同的方法在我们的算法流程中进行测试，以评估它们在从百科网站和社交媒体数据中提取语义信息的能力。在这里，我们简要说明使用的每种文本嵌入方法的原理和特点。

一、语言主题模型

常用的是隐狄利克雷分配模型 (Latent Dirichlet Allocation，LDA[1])，一个文本语料库的生成统计模型，在该模型中，每个文档都可以看作是各种主题的混合体，并且每个主题的特征是对单词的概率分布。LDA 可以表示为三层层次贝叶斯模型，给定一个包含 M 个文档和一个包含 N 个单词的字典的文本语料库，LDA 对一个文档 d 的产生过程定义为：(1) 从狄利克雷分布 $Dirichlet(\alpha)$ 采样 θ。(2) 对 d 中的 N 个单词 w_n 有：从多项式分布 $Multinomial(\theta)$ 采样一个主题 z_n，并从多项式条件概率 $P\left(w_n|z_n,\beta\right)$ 中采样 w_n。其中，其中 θ 是混合比例，从参数 α 的狄利克雷先验中获得，α 和 β 都是语料库参数，在生成语料库的过程中一次采样获得。每个文档都是根据主题 $z_{1:K}$ 和 β 下的单词概率生成的。语料库中文

档 d 的生成概率定义为:

$$P(d|\alpha,\beta)=\int_{\theta}P(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_K}P\left(z_K|\theta\right)P\left(w_n|z_K,\beta\right)\right)d\theta \tag{3.8}$$

LDA 在语料库中学习两组参数：给定主题 $P\left(w|z_{1:K}\right)$ 的单词生成概率与给定文档 $P\left(z_{1:K}|d\right)$ 的主题生成概率。因此每个文档都用主题概率 $z_{1:K}$(假设共有 K 个主题) 与特定主题上的单词概率来表示。新文档可以用学习好的 LDA 模型来表示为一系列主题相关的概率分布（将其投影到主题空间）。

二、单词嵌入

Word To Vector(Word2Vec[44])，是谷歌团队 2013 年提出的无监督学习算法，利用两层前馈神经网络模型进行训练，构建词的分布式语义表示，又被称为 embeddings。设计思路很简单，利用句子中单词间的相关性作为监督信号，通过特定单词的预测任务推动模型训练。训练阶段由两种形式：(1)CBOW(Continuous Bag of Word，连续词袋模型)，用给定词的上下文作为输入去预测该单词。(2)Skip-gram(跳词模型)，用给定词作为输入去预测该单词的上下文。

第一种方法具有扩展性和高效性，常被使用在具体算法实现中。每个单词的特征向量表示存储在模型的隐藏层 (第一层) 参数 $W\in\mathbb{R}^{C\times D}$ 中，C 表示语料库单词大小，D 表示特征维度大小。相关向量经过连接或者求和操作后被用于预测任务。给定一个句子中的训练单词与其对于的特征向量 $w_i\in W, i=1,2,...,T$，模型的优化目标是最大化平均对数概率 P，预测任务由基于 softmax 函数多分类器实现，

$$max\ P=\frac{1}{T}\sum_{t=k}^{T-k}\log p\left(w_t|w_{t-k},\cdots,w_{t+k}\right) \tag{3.9}$$

$$p\left(w_t|w_{t-k},\cdots,w_{t+k}\right)=\frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{3.10}$$

对于每个单词 i，y_i 是经过一次非线性变换后的未归一化对数概率输出，由模型的第二层前馈网络计算获得，其中 U,b 是网络的可学习参数，h 是对来自 W 的单词向量的连接 (concat) 或平均 (average) 操作。

$$y=b+Uh\left(w_{t-k},\cdots,w_{t+k};W\right) \tag{3.11}$$

三、文档嵌入

Document To Vector(Doc2Vec[38])，Word2Vec 的扩展形式，把针对单词特征表示推广到整个段落或句子，学习整个段落的分布式语义表示，即当前单词的预

测由它的上下文单词和所在的段落给出，每个段落有唯一的 id。与 word2vec 类似，给定包含 N 个单词的段落，单词向量由 **W** 内的 $\mathbf{w}_n$ 表示，段落向量由 **D** 内的 $\mathbf{g}_j$ 表示，训练阶段同时更新 $\mathbf{w}_n$ 与 $\mathbf{g}_j$ 参数，满足：

$$\arg\max_{\mathbf{D},\mathbf{W}} \frac{1}{N} \sum_{n=k}^{N-k} \log p\left(\mathbf{w}_n | \mathbf{w}_{n-k}, \cdots \mathbf{w}_{n+k}\right) \tag{3.12}$$

概率预测是通过 softmax 函数产生的，其中 θ 是 softmax 网络的可学习参数，可理解为一个线性变换/矩阵乘法操作。h 是向量的连接或平均操作，融合单词向量与段落向量的信息。

$$p\left(\mathbf{w}_n | \mathbf{w}_{n-k}, \cdots \mathbf{w}_{n+k}\right) = \frac{\exp\left(\theta^T \mathbf{x}_n\right)}{\sum_i \exp\left(\theta^T \mathbf{x}_i\right)} \tag{3.13}$$

$$\mathbf{x}_n = h\left(\mathbf{g}_j, \mathbf{w}_{n-k}, \cdots, \mathbf{w}_{n+k}; \mathbf{D}, \mathbf{W}\right) \tag{3.14}$$

四、预训练模型

我们使用 Bidirectional Encoder Representations from Transformers(BERT[8])，一个大规模的预训练语言模型，在 2018 年被提出并引起业界轰动，被广泛应用于各类自然语言应用中，作为基础网络在下游任务上微调可获得很好的效果。训练阶段也是自监督的，网络的所有层从无标注的文本中利用上下文信息，获取深层双向表示。BERT 结构主要由多个 Transformer-Encoder 模型堆叠而成，Transformer[63] 是一个基于自注意力机制、前馈网络层和残差连接操作的深度神经网络模型，可被用于文本的特征编码，注意函数可以描述为将查询 (query) 和一组键-值 (key-value) 对映射到输出的过程，其中 query、key、value 和输出都是向量，分别表示为 Q, K, V。输出为 value(V) 的加权和，其中分配给每个 value 的权重由 qeury 和相应的 key 点积计算得出，d_k 是 key 向量的维度大小。

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.15}$$

同时引入多头注意力，因为比起单个注意力函数，多次使用 Q, K, V 来完成不同的线性投影更有帮助，允许模型在不同的位置共同关注来自不同表示子空间的信息。因此在模型中平行地完成多个注意力函数的计算，并把所有的输出结果链接在一起。

$$\begin{aligned} \text{MultiHead}\left(Q, K, V\right) &= \text{Concat}(\text{head}_1, \cdots, \text{head}_h)\, W^O \\ \text{where head}_i &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{3.16}$$

BERT 的构造的训练方式主要基于单词的掩盖和句子关联性预测，前者在训练中对随机对 15% 的单词进行掩盖 (mask) 操作，并用剩余的单词去预测；后者把前后相关联的句子对中对一个随机替换为其他句子，让模型去判断句子是否相关。通过这样的预训练方式，模型能关注单词级别与句子级别的语义信息，而 BERT 模型最后一层的第一个单词 (token) 的输出可被用来作为整个句子的特征表示。

第四章　基于自监督算法的模型训练

第一节　训练数据集

本文的多模态自监督学习方法利用图像-文本对语义关联来学习视觉特征。因此需要大规模的多模态图文数据集作支撑，目前与之相关的公开无标注数据集主要来自于维基百科网页 (Wikipedia) 和网络社交媒体 (Instagram, Flickr 等) 平台，我们收集了一些作为训练数据，首先介绍我们使用的多模态数据集的基本情况.

表 4.1　使用数据集的基本统计信息

名称	图片数量	文本数量	文本平均长度	语言	利用率
ImageCLEF	100k	35k	1200+	英文	100%
English-Wikipedia	420M	170M	1200+	英文	30%
Webvision-1.0	240M	240M	10+	多语种	40%
InstaCities1M	100M	100M	20+	多语种	100%

维基百科是一个多语种、基于网络的百科全书项目，目前由超过 4000 万篇文章组成，涉及 299 种不同语言。维基百科文章通常由文本和其他类型的多媒体对象（图像、音频和视频文件）组成，因此可以被视为多模态文档。本文实验利用两个维基百科数据集，ImageCLEF[62] 和 English-Wikipedia[50]，数据集的样本示例如图 4.1所示

1. **ImageCLEF** 由 237, 434 张维基百科图片和相关文章组成，文章分为三种语言（英语、德语和法语）的版本，且至少对应至一个图像和图像描述说明 (caption)。我们只考虑英文文章，过滤小图像（<256 像素）和具有 JPG 以外格式的图像，这样我们训练数据集子集由 100, 785 幅图像和 35, 582 篇独立文章组成 (一篇文章对应多张图像)。
2. **English-Wikipedia** 是 ImageCLEF 数据集的全英文扩展版本，从 290 多个百科全书主题中挑选了 5, 614, 418 篇文章和相关的插图，广泛地覆盖了人类的各类知识，预处理与上文相同，过滤最短边小于 256 像素的图片和少于 50 个单词的文章，最后由 420 万张图片和 170 万篇独立文章组成 (平均每篇文章都有 2.3 幅图片)，然而原始的数据集过于庞大，下载缓慢，因此

我们只取了大约 30% 的数据量使用。

(a) Cat (b) Dog

图 4.1 维基百科网页示例，带多张插图的特定词条 (图中为猫和狗)

网络社交媒体上的多模态数据主要来源于平台用户的贡献，网民在社交平台上分享自己的生活动态，或对特定事件发表观点，每天都会有许多来自不同国家和地区的用户在网络上公开地上传各种主题的图片和文字，因此这些无标注的多媒体数据大量且能免费获取。我们用到了两个公开的多媒体无标注数据集，InstaCities1M[16] 和 WebVision[39]，样本示例如图 4.2所示。

1. **Webvision** 数据集包含 240 多万张从 Flickr 网站和 Google 图片搜索中抓取的图片，并提供了这些图像附带的文本信息（标题、用户标记和描述），我们使用的是 1.0 版本，并且只使用了从 Google 收集的数据 (约数据集的 40%)，也未使用数据集自带的伪标签 (把图片分成了 1000 类)。
2. **InstaCities1M** 的数据是从 Instagram 上搜集的，与全球 10 个人口最多的英语城市相关。它包含每个城市 10 万张图片，总共 100 万张图片，分辨率均为 300x300。该数据集是由近期的社交媒体数据形成的，与图像相关联的文本是由上传者提供的描述和标签 (hashtag)。

各数据集的统计信息如表 4.1所示，可以看出，社交媒体数据集与维基百科数据集数之间的差异较大。前者图片与相关文本是一一对应的，文本是几句话与多个单词 (标签) 组成，属于多语种、杂乱的短文本。后者的图片与文本是多对一的关系，文本是几大段的描述或简介，属于纯英文、文档类型的长文本。因此后续的自监督算法针对数据集各自的特点作出调整，充分挖掘杂乱数据中的潜在信息。

(a) Webvision (b) InstaCities1M

图 4.2 网络社交平台上多媒体数据示例，均是用户生成的

第二节 模型训练细节

对于不同的文本编码器来说，LDA 模型使用普通交叉熵函数优化，Word2vec、Doc2Vec、BERT 模型使用带 sigmord 归一化的二元交叉熵函数或基于度量学习的损失函数。对于 Word2vec 这种提取单词级别特征的模型，整个文档的特征表示由所有单词的特征取平均获得。对于维基百科数据集，由于文本长度的限制，无法使用 BERT 模型提取文本特征，且由于图片与文本存在多对一的关系，因此不使用基于度量学习的损失函数，文本特征维度统一为 40 或 400；而社交媒体数据集则无这些限制，文本特征维度统一为 200 或 400(BERT 模型除外，BERT 基础模型输出层维度是 768)。

使用带动量的随机梯度下降算法 [53] 优化模型，每次迭代批次大小 (batch size) 设置为 128，根据收敛情况共迭代模型 40, 000-60, 000 次，初始学习率为 0.01, 动量参数 (momeentum) 为 0.9，每迭代 20, 000 次学习率乘以 0.1。数据增强使用常见的方式：首先把图片尺寸放缩到 256*256 大小，并随机裁剪 224*224 的区域，再以 50% 的概率进行水平镜像翻转，最后让颜色相关的属性 (亮度、对比度、色调、饱和度) 进行随机抖动。

选择带有 BN 层 (批标准化 [30]) 的标准 AlexNet[35] 网络作为视觉编码器，使用 PyTorch[49] 开源框架搭建模型，在 Nvidia 1080Ti 的 GPU 上训练。LDA、Word2Vec、Doc2Vec 使用 Gensim 库[①]实现训练，BERT 使用 Facebbok 的 Fairseq 库[②]提供的预训练模型 (选取 RoBERTa-base[40] 版本)。

[①]https://radimrehurek.com/gensim/

[②]https://pypi.org/project/fairseq/

第五章　自监督算法性能验证

第一节　验证数据集

我们在不同数据集上做了大量实验来验证算法的有效性，评估此自监督方法学习的视觉特征性能。我们希望该特征具有一定的泛化性和辨别性。使用的数据集包括：PASCAL VOC[11]、ImageNet[7]、Places 205[78]，Multimodal-Wikipedia[54]。

PASCAL VOC 被用于图像分类，目标检测，语义分割任务的数据集，常见的是 2007 年与 2012 年的版本，VOC 2007 包含 9, 963 个图片，共 20 类，被划分为 50% 的训练/验证数据和 50% 的测试数据，VOC 2012 类似，只是图片数量增加到了 2 万多，该数据集可被用于评估自监督学习算法在下游任务的迁移能力。

ImageNet 标准的图像分类任务数据集，始于 2009 年，由李飞飞教授等人建立，共有 1000 种不同类别图片，目前可用的 (公开标签) 有 128 万训练集数据和 5 万验证集数据，常用来评价自监督学习训练的网络的特征表示能力。

Places 205 该数据集共有 250 万张图片，分为 205 类，主要是各种各样的场景图，与 ImageNet 类似，被用于图像分类算法的研究，也可用于评价自监督学习算法性能。

Multimodal-Wikipedia 多模态检索数据集，它由 2, 866 个图像文档对组成，分别分成 2, 173 图片-文本对的训练集和 693 图片-文本对的测试集。每个图片-文本对带有类别语义标签 (共十个)，可用于多模态检索算法性能的评测。

第二节　对比实验

首先该算法进行整体的测试，考虑使用不同数据集，文本编码器，损失函数，找到该算法最优的配置模块。利用 PASCAL VOC 2007 数据集，与之前的验证策略 [22, 48] 保持一致，考虑 AlexNet 模型中间层提取出固定的中间特征 (卷积层 conv4-conv5，全联接层 fc6-fc7)，学习一对多的 SVM[2] 分类器 (使用 Python 的 sklearn 库[①]实现)，在 VOC07 的训练集上训练该分类器，根据在验证集的结果调整超参数，最后在测试集上计算多分类的评价指标 (mAP)。

文本编码器　表5.1(a)展示基于维基百科数据集的结果，使用交叉熵损失，括

[①]https://scikit-learn.org/stable/

号内为输出特征维度大小，可以看出使用 LDA 模型提取文本特征的效果最好，由于该数据集的文本都是大篇段落，数据结构性强，不同文本之间语义差异性明显，用基于主题模型的算法能提取出深层次的语义信息，适合作为监督信息来约束视觉网络学习相似的表征。表5.1(b)展示基于社交媒体数据集的结果，可以看出 Word2Vec 模型效果最好，并且与维基百科数据集效果差了很多。这是由于该数据集每张图片对应的文本长度较小，且噪声更大，存在很多非语言字符，又是多语言版本，基于文本级别的语言模型无法很好的发挥作用，即使是 BERT 模型也未有改善。文本语义信息太少，很难提取到对应的特征，因此文本监督信号太弱了，无法指导视觉网络进行有效的学习。

数据集种类与大小 表5.1(c)展示不同数据集结合与其对应最优的文本编码器的比较，由于 Wikipedia 是 ImageCLEF 的数据扩充版本，可以看出增大数据集的大小，能适当提升算法的性能，这在机器学习中是很常见的结论，数据越多，模型能学习到的样本种类越丰富，输出的特征表示性越强。并且数据的质量也会影响最后的结果，WebV(Webvision) 数据集对我们的算法更友好一些，可能因为文本噪声比 Ins1M(InstaCities1M) 更少。

损失函数 表5.1(d)展示不同损失函数比较，使用 InstaCities1M 和 Word2vec 文本编码器。可以看出对特征不做任何处理，直接使用均方误差损失函数 (MSE) 优化效果最差，同时观察到在训练过程中目标函数不收敛。使用 Sigmoid 与二元交叉熵损失函数 (BCE) 优化时效果会有较大的改善，并且当使用度量学习常用的损失函数 (triplet, contrastive) 时，性能进一步的提升。

总的来说，选择维基百科数据集和 LDA 文本编码器，在我们的多模态自监督算法框架上表现最好，并且使用的无标注数据越多，在下游任务表现的性能越突出，这符合我们的预期。同时对于社交媒体数据集，使用 word2vec 提取文本特征，对 CNN 输出的视觉特征 L2 归一化，并用度量学习的损失函数在统一的度量空间内进行优化，效果最好。最后我们使用不同种类数据集对于的最优参数组合，与之前的自监督学习方法进行比较，结果如表 5.2所示，可以发现我们的自监督算法超过了 2016 年之前主流算法，但与全监督的算法差距还是比较大。

第三节 线性分类器

与之前的研究工作保持一致，我们固定网络中间层的特征，在 Imagenet 和 Places 数据集上学习线性分类器。因为线性分类器的分辨能力很弱，因此十分依

表 5.1　对比实验的结果

(a) ImageCLEF 与不同文本编码器

Method	conv5	fc6
LDA(40)	**47.1**	**48.3**
Word2Vec(40)	43.5	43.8
Word2Vec(400)	40.4	44.9
Doc2Vec(40)	41.2	39.5
Doc2Vec(400)	42.6	34.5

(b) Ins1M 与不同文本编码器

Method	conv5	fc6
LDA(200)	34.3	30.2
LDA(400)	36.5	32.7
Word2Vec(200)	**40.2**	**36.5**
Doc2Vec(200)	32.4	26.9
BERT(768)	35.1	32.4

(c) 使用不同类型与规模数据集

Dataset	Type	conv5
ImageCLEF+LDA	Wiki	47.1
Wikipedia+LDA	Wiki	**50.8**
Ins1M+Worc2Vec	Social	40.2
WebV+Worc2Vec	Social	41.3

(d) 使用不同的损失函数

Loss func	Norm	conv5
MSE loss	None	30.1
BCE loss	sigmoid	40.2
triplet loss	L2-norm	41.4
contrastive loss	L2-norm	**42.7**

赖于输入特征的好坏。使用 AlexNet 模型，提取每个卷积单元 (conv1-conv5) 经过激活函数后 (ReLU) 输出的特征，评估策略参考 [3, 73], 因为每层原始输出的特征维度不同，为了保证公平性，对其经过池化层下采样到维度差不多的大小，测试时对每一张图片进行十次裁剪，并计算十张图的平均准确率作为该样本图片的准确率。ImageNet 结果如表 5.3，Places 结果如表 5.4所示。可以看出我们的方法能达到 2018 年的基准 (baseline) 水平，超过大部分单模态的自监督算法，减小了与 ImageNet 全监督预训练模型的差距。

第四节　迁移学习

验证自监督算法特征的迁移学习能力，依然选择 PASCAL VOC 数据集，该数据集每个类别的样本数很少，更接近真实世界的情况。把自监督训练好的网络作为模型权重的初始化，在下游任务 (多标签分类、目标检测、语义分割) 上进行微调，分类与检测基于 VOC07 数据集，分割基于 VOC12 数据集。AlexNet 作为骨干网络，目标检测选用 Fast-RCNN 框架 [14]，语义分割选用 FCN 框架 [41]。评估策略与代码参考 [3, 34]，迭代训练 80, 000 次，使用 SGD+momentum 优化

表 5.2 不同算法在 VOC 2007 分类任务上的性能，在 AlexNet 不同层的固定特征上训练 SVM 分类器，在测试集计算的% mAP 得到的结果。

Method	Ref	conv4	conv5	fc6	fc7
ImageNet Supervised(2012)	[15]	-	65.6	69.6	73.6
Places Supervised(2014)	[15]	-	63.2	65.3	66.2
Sound[48](2016)	[48]	-	46.7	47.1	47.4
Tracking[68](2015)	[68]	-	42.2	42.4	40.2
Jigsaw[46](2016)	[22]	**53.0**	-	-	-
Colorization[75](2016)	[22]	49.0	-	-	-
Ours(Ins1M+Word2Vec)	-	42.4	42.7	41.5	40.1
Ours(Wikipedia+LDA)	-	51.2	**50.8**	**53.2**	**54.0**

表 5.3 不同算法在 ImageNet 数据集上的结果 (验证集的 top-1 准确率%)，使用 AlexNet+ 线性分类器，训练时冻结卷积层，* 表示使用更大版本的 AlexNet 网络结构

Method	conv1	conv2	conv3	conv4	conv5
ImageNet Supervised	19.3	36.3	44.2	48.3	50.5
Random	11.6	17.1	16.9	16.3	14.1
Jigsaw(2016)	18.2	28.8	34.0	33.9	27.1
Colorization(2016)	12.5	24.5	30.4	31.5	30.3
SplitBrain[76](2017)	17.7	29.3	35.4	35.2	32.8
Rotation[13](2018)	18.8	31.7	**38.7**	38.2	**36.5**
DeepCluster*[3] (2018)	12.9	29.2	38.2	39.8	36.1
Ours(Ins1M+Word2Vec)	16.4	26.5	31.0	29.3	26.3
Ours(ImageCLEF+LDA)	18.1	29.5	35.1	32.0	29.2
Ours(Wikipedia+LDA)	**20.0**	**32.1**	37.3	**38.6**	33.4

器，在验证集上调整超参数，在测试集展示结果，如表 5.5所示。可以看出算法学习到的特征在迁移学习任务上表现出色，能与近几年主流的自监督方法相竞争，并且在目标检测的任务上性能最好，充分证明了特征的强大的泛化能力。

表 5.4　不同算法在 Places 205 数据集上的结果 (验证集的 top-1 准确率%)，使用 AlexNet+ 线性分类器，训练时冻结卷积层，* 表示使用更大版本的 AlexNet 网络结构

Method	conv1	conv2	conv3	conv4	conv5
Places Supervised	22.1	35.1	40.2	43.3	44.6
ImageNet Supervised	22.7	34.8	38.4	39.4	38.7
Random	15.7	20.3	19.8	19.1	17.5
Jigsaw(2016)	**23.0**	32.1	35.5	**34.8**	31.3
Colorization(2016)	16.0	25.7	29.6	30.3	29.7
SplitBrain(2017)	21.3	30.7	34.0	34.1	32.5
Rotation(2018)	21.5	31.0	35.5	34.6	**33.7**
DeepCluster*(2018)	18.6	30.8	37.0	37.5	33.1
Ours(Ins1M+Word2Vec)	16.4	25.4	28.9	28.1	27.2
Ours(ImageCLEF+LDA)	19.8	28.7	31.4	30.1	29.5
Ours(Wikipedia+LDA)	21.6	**32.4**	**35.6**	34.0	32.7

第五节　多模态检索

作为多模态场景的特征学习算法，评估学习到的视觉特征在多模态检索任务上的性能。我们使用 MultiModal-Wikipedia[54] 数据集，分为图片查询文本 (Image Query) 与文本查询图片 (Text Query) 两种任务形式，分别提取图片和文本的特征，并求得查询对象 (图片或文本) 与数据库中所有数据的距离 (相对熵或余弦距离)，按相似度从大到小排序并计算查询 mAP(mean Average Precision)，结果如表 5.6所示。

表中所有的跨模态检索方法都使用对应的文本编码器 (语言模型) 提取文本特征，视觉编码器 (卷积神经网络，CNN) 提取图像特征，但与自监督学习方法相比，表中的有监督学习的多模态检索方法利用了图片-文本对的类别标签信息，也使用了有监督训练的 ImageNet 预训练模型作为网络初始化。可以看出我们的自监督算法在跨模态检索任务上具有一定的出色表现，并且检索性能与有监督的方法差距很小。

表 5.5 不同的自监督学习算法，在 PASCAL VOC 数据集上进行迁移学习的结果 (图像分类、目标检测、语义分割任务)

Method	Classification (%mAP)	Detection (%mAP)	Segmentation (%mIoU)
ImageNet Supervised	79.9	59.1	48.0
Random	53.3	43.4	19.8
Jigsaw(2016)	67.6	53.2	37.6
Colorization(2016)	65.9	46.9	35.6
SplitBrain(2017)	67.1	46.7	36.1
Rotation(2018)	**73.0**	54.4	**39.1**
DeepCluster*(2018)	74.7	55.4	45.1
Ours(ImageCLEF+LDA)	66.9	50.3	36.7
Ours(Wikipedia+LDA)	70.3	**54.6**	38.5

表 5.6 在 Multimodal-Wikipedia 数据集上不同算法的跨模态检索结果 (%mAP)，带 * 表示有监督方法，使用了类别信息

Method	Image Query	Text Query	Average
CCA[24](2010)	19.7	17.8	18.8
PLS[55](2006)	30.6	28.0	29.3
JFSSL*[65](2016)	**42.8**	39.6	**41.2**
CCA-3V*[18](2014)	40.5	36.5	38.5
LCFS*[66](2013)	41.9	38.5	39.9
Ours(Ins1M+Word2Vec)	32.2	30.6	31.4
Ours(ImageCLEF+LDA)	36.8	36.2	36.5
Ours(Wikipedia+LDA)	37.4	**39.8**	38.6

第六章　结论

在这篇文章中，我们提出了一种基于多模态场景的自监督学习算法，利用互联网上大量可获取的无人工标注多模态数据，构建视觉的表示学习算法。通过文本与图像之间的内容相关性，用文本的语义信息作为监督信号，指导卷积神经网络训练，学习到一定的特征表示能力。我们尝试了不同类型的无标注的多模态数据集，用不同语言模型作为文本编码器提取文本特征，并结合了度量学习领域的多种损失函数，基于此做了细致的消融与对比实验，找到了最适合我们算法框架的参数配置与模块组合。为了验证自监督算法学习特征的有效性，最后我们在图像分类，目标检测，跨模态检索等计算机视觉任务中验证该自监督方法学习到的表示特征，发现其能很好地迁移到各种下游任务中，特征的分辨能力与泛化能力超过大部分单模态的自监督学习算法，在各种评测指标上能得到较好的结果。

参 考 文 献

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. *Lecture Notes in Computer Science*, page 139–156, 2018.

[4] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2959–2968, 2019.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[6] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12154–12163, 2019.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[10] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[12] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[15] Lluis Gomez, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4230–4239, 2017.

[16] Raul Gomez, Lluis Gomez, Jaume Gibert, and Dimosthenis Karatzas. Learning to learn from web data through deep semantic embeddings, 2018.

[17] Raul Gomez, Lluis Gomez, Jaume Gibert, and Dimosthenis Karatzas. Self-supervised learning from web data for multimodal retrieval, 2019.

[18] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233, 2014.

[19] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European conference on computer vision*, pages 529–545. Springer, 2014.

[20] Albert Gordo, Jon Almazán, Naila Murray, and Florent Perronin. Lewis: latent embeddings for word images and their semantics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1242–1250, 2015.

[21] Albert Gordo and Diane Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6589–6598, 2017.

[22] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6391–6400, 2019.

[23] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[24] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[28] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[29] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding, 2019.

[30] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[31] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[32] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision*

and pattern recognition, pages 3128–3137, 2015.

[33] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

[34] Philipp Krähenbühl, Carl Doersch, Jeff Donahue, and Trevor Darrell. Data-dependent initializations of convolutional neural networks. *arXiv preprint arXiv:1511.06856*, 2015.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[36] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[37] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision*, pages 577–593. Springer, 2016.

[38] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[39] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data, 2017.

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[41] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[42] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining, 2018.

[43] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watch-

ing hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2630–2640, 2019.

[44] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[45] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[46] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[47] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[48] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[50] Yash Patel, Lluis Gomez, Raul Gomez, Marçal Rusiñol, Dimosthenis Karatzas, and C. V. Jawahar. Texttopicnet - self-supervised learning of visual features through embedding images on semantic text spaces, 2018.

[51] Yash Patel, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. Dynamic lexicon generation for natural scene images. In *European Conference on Computer Vision*, pages 395–410. Springer, 2016.

[52] Yash Patel, Lluis Gomez, Marçal Rusiñol, Dimosthenis Karatzas, and C. V. Jawahar. Self-supervised visual representations for cross-modal retrieval, 2019.

[53] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[54] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal

multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260, 2010.

[55] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pages 34–51. Springer, 2005.

[56] Mohammad Sabokrou, Mohammad Khalooei, and Ehsan Adeli. Self-supervised representation learning via neighborhood-relational encoding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8010–8019, 2019.

[57] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3020–3028, 2017.

[58] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[59] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning?, 2019.

[60] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019.

[61] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2019.

[62] Theodora Tsikrika, Adrian Popescu, and Jana Kludas. Overview of the wikipedia image retrieval task at imageclef 2011. In *CLEF (Notebook Papers/Labs/Workshop)*, volume 4, page 5, 2011.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[64] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference*

on Multimedia, pages 154–162, 2017.

[65] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2010–2023, 2015.

[66] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. Learning coupled feature spaces for cross-modal matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2088–2095, 2013.

[67] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[68] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[69] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[70] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[71] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification, 2019.

[72] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. *arXiv preprint arXiv:1912.03330*, 2019.

[73] Asano YM., Rupprecht C., and Vedaldi A. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020.

[74] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning, 2019.

[75] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[76] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Un-

supervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

[77] Junbo Zhao, Michael Mathieu, Ross Goroshin, and Yann Lecun. Stacked what-where auto-encoders. *arXiv preprint arXiv:1506.02351*, 2015.

[78] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.