中国科学技术大学

University of Science and Technology of China

本 科 毕 业 论 文

题　目 视觉可解释的时尚服装特征表示

英　文 **Visually Explainable Feature Representations**

题　目 **of Fashion Apparel**

院　系 计算机科学与技术学院

姓　名 史恒锐　　学　号 **PB17111565**

导　师 刘 淇

日　期 **2021.5.8**

致　谢

毕业来临之际，不由得感慨万千，时光荏苒，光阴如梭，本科四年转瞬即逝。在整个本科生活里，我认识了很多人，也遇到了很多事。来自五湖四海的同学让我放下了心中离家千里的不安，老师教授们也能充分地指导我的学业。首先，我要感谢我的导师，刘淇教授。没有刘淇老师的悉心教导，我很难完成这个项目，老师的耐心、宽容和丰富的知识、经验让我稳步前进，脚踏实地。我很感谢老师能给我完成此次研究的机会，他教会了我如何完成研究。老师的真挚和见解不断激励着我，我也很珍惜这段终生难有的时光和经历。同时，我还要感谢我的师兄，在老师事务繁忙的时候能向我伸出援手，回答我的疑问，帮助我克服遇到的困难，还能在学习生活上给我建议，老师指引我的方向的同时，师兄也会在细节上给予我更多的支持。

其次，我要感谢曾经帮助过我的同学们。无论是生活还是学业，我都受到过许多人的热心帮助，即使是一件件微不足道的小事，也能让我感受到整个院系、整个班级的联结。或是教会我如何解题，或是帮助我解决研究过程中遇到的问题，每一次的帮助我都会铭记于心，在别人需要的时候我也会尽我所能。不论是举手之劳还是无心之举，任何帮助都有可能救人于水火，因此我再次感谢那些愿意向我提供帮助的同学。

我还要感谢学院和学校，学院设立的课程让我得以在这个专业领域正是迈出第一步，为我以后的学术和职业生涯奠定了牢固的基础。虽然有时会显得学业繁重，但是现在回过头看，也物有所值，学有所成，坚实的数理基础让我能比更多人有更强的竞争力。而学校提供的平台让我能够接触到更广阔的天地，让我见识了优秀的人才和之前从未体验过的事物。本科的四年让我开阔了眼界，这是我一生中难忘的四年。

最后，我要感谢我的父母，没有人比我父母更关心我的学习和生活，我这四年少不了他们的支持和鼓励，只身在外，只有父母才能给予我精神上的慰藉。有了他们的支持，我才能顺利完成我的学业，才能完成我的毕业设计。虽然分隔两地，我依然能感受到父母的心意。谨以此，送给所有帮助我的人。

目 录

中文内容摘要

在深度学习中，数据对象的特征是非常重要的，深度模型利用并学习这些特征，最后生成我们感兴趣的输出。在这个过程中，如何表示特征是很多研究中共同关注的部分，是否能够有效地表示特征，决定了模型是否能达到较高的性能。因此，找到一个优秀的方法提取并表示数据中的特征至关重要，模型将学习到特征中独特的部分，来适应特定的对象以输出理想的结果。

现实生活中，任何物品都有复杂的属性构成，对于衣物而言，如袖长、领口、衣长等一起表现出了一件衣服的特点。不同的人可能喜欢不同的衣服特点，比如某人更喜欢 V 领而不是圆领，喜欢穿宽松的衣服而非贴身的。在服装店里，每个店员都会询问这些信息，确定了这些喜好就能更好地找出最有可能令客人满意的商品，但是在大多数的研究中，深度学习模型仅仅关注了服装图像的整体情况，而忽视了这些属性的信息，从而很难产生较好的特征表示。

因此，在这篇论文中完成了一个专注于物品（服装）各类属性的模型，可将服装的各类特征投影到某个空间中，根据不同特征的标签信息，模型能够在服饰图像上找到对应的关注点，在提供物品表征信息的同时，还能从视觉上展示目标特征的重要性。这些特征信息可以被用在其他领域中，例如图像分类和推荐。

Abstract

In deep learning, the features of data objects are very important, which are utilized and learned by deep learning models to finally generate the output we are interested in. In this process, feature representation is a common part of concern in many studies. Effective and accurate feature representations determine whether the model is able to achieve high performance. Therefore, it is crucial to find an excellent way to extract and represent the features in the data, with which the model will learn the unique parts of the data to adapt to the specific objects to come to desired results.

In real world, any item has a complex composition of attributes, which for clothing, such as sleeve length, collar design, and garment length, together represent the characteristics of a piece of clothing. Different people may prefer different characteristics of clothes, for instance, someone prefers V-neck rather than round neck, and the other prefers to wear loose fitting clothes rather than fitted ones. In a clothing store, it is very normal and habitual for each clerk to ask for this information, and determine if these preferences will better identify the items that are most likely to satisfy the customer. Nevertheless, deep learning models in most studies focus only on the overall pictures of the clothing images but ignore vital, necessary information about these attributes, making it difficult to produce better feature representations.

Therefore, a model concentrating on various types of attributes of items is requested, which is completed in this thesis. Different types of features of clothing are independently projected into a certain space in this model. Based upon the labeling information of different features, the model is able to find the corresponding focal points on the clothing images, and provide the item representation information as well as show users the significance of those highlighted features visually. This information of features can be well used in other related fields, such as image classification and recommender system. As mentioned above, if a garment is a loose fitting one, then the recommender system fed with feature representations we got “believes” the user will be interested in the item. In this way, this recommender system is capable of collecting more reasonable items as candidates to improve its performance.

第一章　绪论

第一节 背景

抽取数据特征并生成对应的表示在现有的许多研究中是司空见惯的，所有深度学习模型都需要对原始数据进行特征提取和表示，以便模型学习这些特征输出结果并做出决策。然而，以何种方式表示这些特征却尚无一个标准，不同的数据有不同的特征表示方法，不同的数据侧重的特征也可能不同，特征表示的质量高低关系到模型性能的好坏，不够准确的特征表示可能导致模型对输入数据毫无作为，不能产生任何有实际意义的结果。因此，研究如何表示数据中的特征是十分有必要的。

此外，特征表示的可解释性也是关键的问题之一。由于抽象的数据难以和现实中的对象一一对应，大多数深度学习模型所学习的特征形式对人类来说是不可解释的，也就是说人类无法理解这些特征有何实际意义、为何模型能从这些特征中学习到某些特殊信息，这些困难阻碍着人们理解他们的模型，阻碍模型的优化进一步发展。所以在本文中，将从服装图像的角度出发，研究如何用人类可以理解的过程来生成这些图片的特征表示，最终可以看到，这些特征都以人类视觉上可解释的形式呈现，任何人都可以轻松得知这些特征对应的是服装的哪一特点。利用这种特征表示方法，这些信息即可进一步输入到其他模型中，比如推荐系统，从而实现更加精确的推荐。

接下来，将介绍一些技术背景和相关研究，在相关工作中将会见到一些处理特征的方式和在其他如推荐系统领域中的应用。

一、数据特征

1. 概念

在机器学习相关领域中，数据向来是让人重视的部分，没有质量上乘的数据是不能得到性能优异的模型的，而模型输出的效果和数据的特征表示息息相关，如果这些数据表征能清晰地体现出原始数据中的特性，那么模型会成功学习到这些特性，作为输出结果之前的关键之处。

在计算机视觉和图像处理中，特征是关于图像内容的一段信息，通常是关于图像的某个区域是否具有某些属性。特征有可能是图像中的特定结构，比如点、边缘或物体，也可以是应用于图像的一般性区域操作或特征检测的结果。同时还

存在其他类型的特征，例如图像序列中移动的情况，或者是不同图像区域之间的边界和曲线。

更一般地，特征是任何与解决与某一应用相关的计算任务有关的信息，比如人脸识别中瞳距、眉毛形状、颧骨高度等等。尽管大多数图像处理都有一个非常复杂的特征集合，但这些特征与机器学习和模式识别中的特征意义相同。特征的概念实际上是非常普遍的,在一个特定的计算机或者视觉系统中对于特征进行选取很有可能是高度地依赖于所需要解决和处理的具体问题,因此对于什么是特征，并没有一个普遍或确切的定义，确切的定义往往取决于问题或应用的类型。尽管如此，一个特征通常被视作图像中 “令人感兴趣”的部分，而且特征被当作许多计算机视觉算法的基础，由于其是后续算法的关键，整个算法往往只能表现出一个特征检测器那样的性能,所以一个理想的特征检测器能够在同一场景下的两个或多个不同图像中检测到相同的特征。

2. 特征检测

特征检测是一个十分基础的图像处理操作,通常作为图像处理的第一个步骤来执行，检查每个像素是否有特征存在。如果这是某个较大的算法的一部分，那么该算法通常只检查特征区域内的图像。特征检测前提是，输入图像通常由高斯核在标量空间表示中进行平滑处理，再计算一个或几个特征图像，常用的表示为局部图像的求导。有些时候,当特征检测的计算成本较高且对时间有较高要求时，可使用更高级别的算法来实现，亦即只对图像的某些部分进行特征检测。

特征检测包括计算图像信息的抽象表示，并在每一个图像点上做出判断，在该点上是否存在一个给定类型的图像特征。由此产生的特征将是图像的子集，通常以孤立的点、连续的曲线或连接的区域的形式出现。有名的特征检测器包括最大稳定极值区域（MSER）[1]。

3. 特征提取

在机器学习、模式识别和图像处理中，从原始数据集开始进行特征提取，剔除重复度较高的特征,计算非冗余的特征属性,以便进行后续的特征学习和归纳，再生成理想的结果，在某些情况下使得人类可以更好地解释这些特征。特征提取与降维有关[2]。

当一个算法的输入数据太大，无法处理，并且它的成分较为冗余（例如，用毫米或者厘米表示相同的衣服长度，或者是服装图像背景的重复性），那么它可以被转化为一个更小的特征集（也被称为特征向量）。确定初始特征的一个子集

被称为特征选择，所选择的特征可能包含输入数据中的相关信息，这样就可以通过使用这个精简的表示方法而不是完整的原始数据来执行所需的任务。

特征提取的目的是减少描述一大组数据所需的资源数量。在对复杂的数据进行分析时，一个主要的问题来自于所涉及到的变量数量。有大量变量的分析通常需要消耗大量的内存和计算能力，同时也可能导致分类算法对训练样本过拟合，对新样本的泛化能力很差。特征提取是构建变量集合的方法的总称，这些方法是为了绕过这些问题，同时还能足够准确地描述数据而提出的。特征工程是为了改善结果而构造与实际应用相关的特征集合，比如图像处理。

4. 特征表示

一个特定的图像特征，以图像数据中的特定结构来定义，通常可以用不同的方式表示。例如，一个边缘可以被表示为每个图像点的布尔变量，描述该点是否存在一个边缘。另外，我们也可以使用一种表示方法将其与关于边缘方向的信息结合起来，提供一种确定性的衡量标准，而不是对边缘存在的布尔式声明。同样地，一个特定区域的颜色可以用三个颜色通道（RGB）来表示。

在某些应用中，只提取一种类型的特征不足以从图像数据中获得相关信息，需要提取两种或多种不同的特征，从而在每个图像点上产生两种或多种特征描述。一个常见的做法是将所有这些特征提供的信息表示成一个单一向量的元素，这个向量通常被称为特征向量。所有可能的特征向量的集合构成了一个特征空间。一个常见的特征向量的例子是，每个图像点要被分类为一个特定的类别。假设每个图像点都有一个基于合适的特征集的相应的特征向量，也就是说，每个类在相应的特征空间中被很好地分隔开，那么每个图像点的分类就可以用标准的分类方法完成。

另一个相关的例子是基于神经网络的图像处理。神经网络的输入数据通常是以每个图像点的特征向量的形式表示的，其中特征向量是由从图像数据中提取的几个不同的特征构成，网络本身可以学习哪些不同特征的组合对解决指定问题是有用的。

当设计一个计算机视觉系统或计算机视觉算法时，特征表示的选择可能是一个关键问题。在某些情况下，为了解决问题，可能需要更高层次的特征描述，但这是以必须处理更多的数据和更苛刻的处理为代价的，通常需要更复杂的模型和更大的时间成本。

二、深度学习

1. 简介

深度学习是一种机器学习算法，深度学习模型利用多个网络层处理输入，逐步提取更深层次的特征获得输出，每层网络都会学习如何抽象表示输入数据。简单来说，“深度”的意思就是网络的层数，深度为 2 的网络是一个通用的近似工具，可以模拟任何函数[3]，虽然增加网络深度并不能提高网络模拟函数的近似程度，但是深度模型比浅度模型能提取更好的特征，从而产生更好的学习能力。

深度学习算法，如深度置信网络，可以用于无监督学习。同时，深度模型通过将复杂的原始数据转化为类似主成分的表征，消除数据中的冗余，因此能够取消特征工程，在监督学习中也有优秀的表现。

2. 卷积神经网络

卷积神经网络来源于神经元的功能和过程，单个神经元只会处理有限区域的信号，不同神经元的感受野有部分重叠，从整体上看所有神经元覆盖了全部区域。和其他神经网络类似，卷积网络同样由输入、输出、权重、激活函数等组成，可以模拟复杂的非线性关系，已经证明了稀疏多元多项式用深度网络模拟比浅层网络更容易[4]。在许多不同的深度神经网络中，卷积神经网络是其中最为广泛使用的网络之一。

卷积神经网络常常被用于图像分析，如今也在自然语言处理、推荐系统等各个领域中大量使用。卷积神经网络由全连接网络改进而来，全连接网络容易导致过拟合，常见的避免过拟合的方法有权重衰减、正则化等，而卷积网络则利用数据中的某些层级模式，通过过滤器（卷积核）的较为简单的模式组成更大更复杂的模式，因此卷积网络的复杂性通常都很小。

与其他图像分类模型相比，卷积网络的预处理过程较少，相对于人工干预特征提取的先验知识，卷积网络通过学习来自动优化过滤器，这是一个独特的优势。

三、可解释的人工智能

可解释的人工智能（XAI）是指解决方案的结果可以被人类理解的人工智能，它与机器学习中的“黑匣子”概念形成鲜明对比，在机器学习中，即使其设计者也无法解释为什么人工智能会做出某些决定。XAI 可以通过帮助终端用户相信人工智能正在做出正确的决定来改善产品或服务的用户体验，这样一来，XAI 的目

的是解释已经做了什么，现在做了什么，接下来会做什么，并揭开行动所依据的信息[5]。这些特点使其有可能确认现有知识、挑战现有知识和产生新的假设。

人工智能中使用的算法可以被区分为白盒和黑盒机器学习算法。白盒模型提供的结果对该领域的专家来说是可以理解的,而黑盒模型则是极难解释的[6]。XAI 算法被认为要遵循透明、可解释和可说明的三个原则。透明度是指如果从训练数据中提取模型参数和从测试数据中生成标签的过程可以被描述清楚[7]。可解释性描述了理解模型的可能性,并以人类可以理解的方式呈现决策的基本依据[8][9]。可解释性是一个被认为很重要的概念，但目前还没有一个共同的定义。有人提出，可解释性可以被认为是〞可解释领域的特征的集合，这些特征对一个给定的例子产生决策[10]，例如分类或回归。如果算法满足这些要求，它们就为证明决策、跟踪从而验证决策、改进算法和探索新事实提供了基础。

有时也可以用本身可解释的白盒算法实现高准确度的结果。这在医学、国防、金融和法律等领域尤为重要，因为在这些领域，理解决策并建立对算法的信任是至关重要的[5]。

算法和人类之间的合作，取决于信任。如果人类要接受算法的决策，他们需要信任它们。人工智能系统有时会学习到不够理想的技巧，这些技巧在训练数据上做了满足明确的目标的最佳工作,但并不能反映系统设计者的复杂的初始愿望。例如，2017 年一个负责图像识别的系统学会了“作弊”，它寻找一个恰好与马匹图片相关的版权标签，而不是学习如何判断图片中是否真的有一匹马。在另一个 2017 年的系统中，一个负责在虚拟世界中抓取物品的监督学习人工智能也学会了“作弊”,它将其操作工具放在物体和观众之间,使其看上去正在抓取该物体。从结果上看，系统似乎能够解决初始任务，但整个过程实际上是非常糟糕的。

而在本文中，将展示一个能够生成可解释特征提取和表示的模型，这些特征信息当被用于其他领域时,能够使最终的输出结果有较高的可信度和较好的可解释性，比如在推荐系统中，用户可以通过系统看到为何会推荐某件服装，是因为它是圆领还是因为它是长袖。

第二节 相关工作

一、定位特征

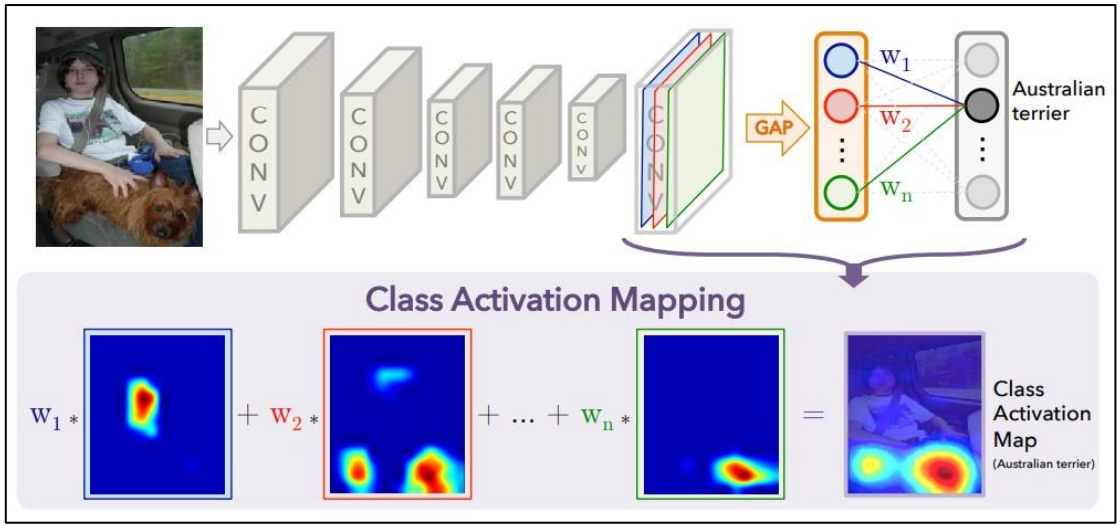

图 1 CAM 结构[11]

语义属性作为理解图像的细粒度表征，在计算机视觉领域中是一类重要的信息。类别激活图[11]（Class Activation Maps，后文简称 CAM）是一种能够标明与类别信息最相关的图像区域的方法，CAM 将卷积网络生成的特征图进行全局平局池化，和模型输出加权求和，得到最后的激活图，原理如图 1 所示。在 CAM 的基础上，发展出另一种相关的技术 AAM[12]（Attribute Activation Maps），AAM 是一种以弱监督方式定位并表示属性的方法，如图 2 所示，在全局平均池化后将得到的 ROI 信息和卷积层中的特征图结合，通过 ROI 池化的方式生成对应的特征向量。已经有研究表明，利用局部表征方法对属性建模是一种更有效的方式[13]。

从图 1 可以看出，CAM 将网络中的全连接层替换为全局平均池化层（GAP），这导致需要修改模型并重新训练，开销不容小觑，为了简化这一过程，Grad-CAM

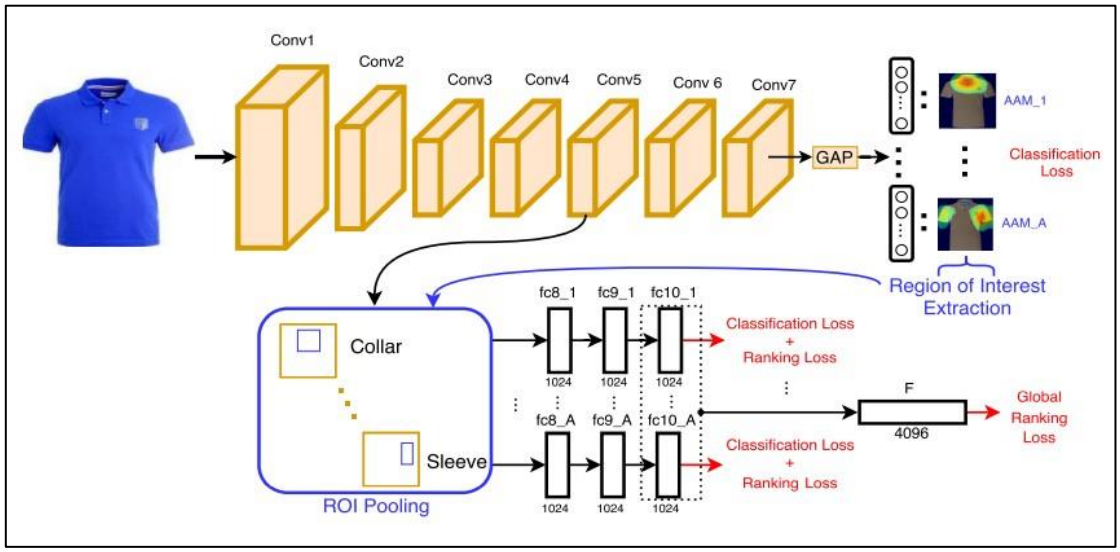

图 2 AAM 结构[12]

应运而生[14]。CAM 利用 GAP 求出权重，而 Grad-CAM 仅需网络梯度信息的全局平均即可求得权值，其原理如图 3，两者求得的权重在数学上是等价的。除了 Grad-CAM 生成的特征图，模型中的反向传播也可用来解释分类结果，其分类效果不会比 CAM 差，但在训练成本上有极大的提升。在图 4 中可以看到 Grad-CAM 的效果。

二、可解释的推荐系统

传统的推荐系统缺乏可解释性，如前文所述，大部分模型会将图像作为一个整体处理，因此很难给出令人信服的推荐理由。目前已经有越来越多的研究专注于推荐的可解释性，将矩阵分解中的显性因素作为可解释的部分，加强了协同过滤算法的可解释性[15]，这些研究将矩阵隐变量与特定物品的显式主题一一对应，来提高模型的说服力。

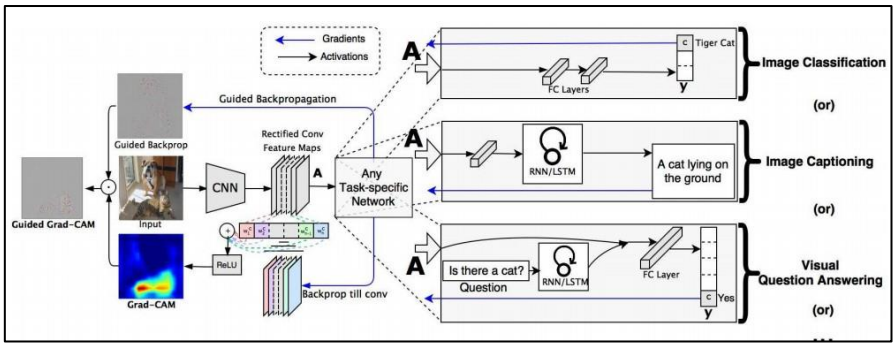

图 3 Grad-CAM 结构[14]

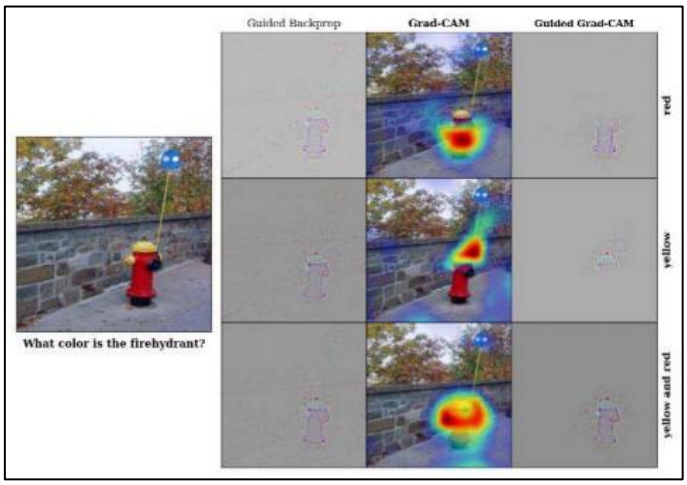

图 4 Grad-CAM 效果[14]

对于深度神经网络，也有研究发现注意力机制可以自主学习物品各种显性特征的重要性并改善用户和物品的表示[16]。但这些研究仅停留在物品层面。

三、推荐的视觉效果

视觉推荐利用数据中的视觉信号对物品和用户偏好的视觉特性进行建模。一些工作如[17]，利用在 ImageNet 上预训练的卷积网络直接生成物品的视觉表示。另一些研究[18]去除了物品的分类特征，从卷积网络生成的视觉特征向量中提取了物品的风格特征来表示物品。此外，另有一种美学网络用来获取图像的美学元素以完成时尚推荐[19]。以上这些方法注重物品的整体特性，将物品图像用一个固定的隐向量表示，因此物品的细粒度属性并未得到充分利用。

而在目标检测中，较为出色的研究有 R-CNN 系列。首先来关注 Fast R-CNN，其结构图如图 5 所示。Fast R-CNN 从图像中找出 ROI（Region of Interest），通过

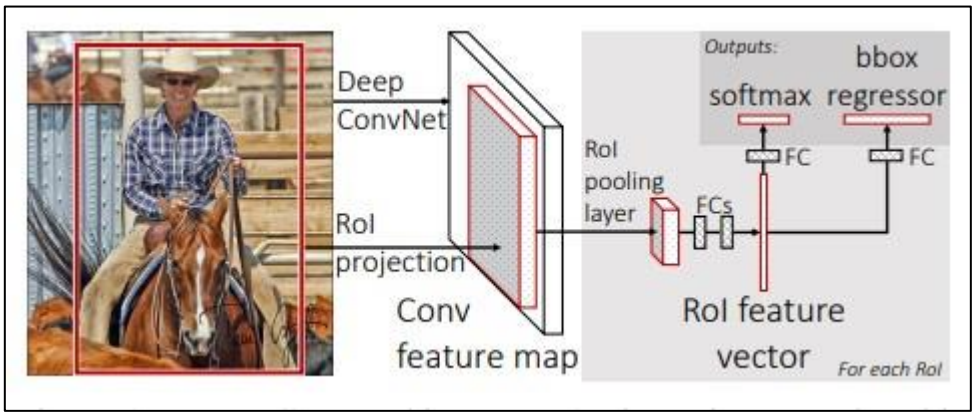

图 5 Fast R-CNN 结构[20]

卷积网络和 ROI 投影后得到对应的特征图，再经过 ROI 池化层获得 ROI 特征向量，经过全连接层后求得图像的 softmax 分类结果和 bounding box 信息，bounding box 如图 5 中图片里的红框所示。ROI 池化的好处是，即是 bounding box 的尺寸大小不尽相同，在池化处理后能得到相同尺寸的输出，ROI 池化层可以接受任意尺寸的输入，这相比于之前的池化层有非常大的改进。Faster R-CNN 由 Fast R-CNN 改进而来[21]，作者在卷积层后设计了一个 RPN 结构（Region Proposal Network），用来提取图像中的所有 bounding box，这个结果称为 proposals，再将 proposals 和特征图结合输入 ROI 池化层，通过分类器即得分类结果。Faster R-CNN 的结构如图 6 所示。

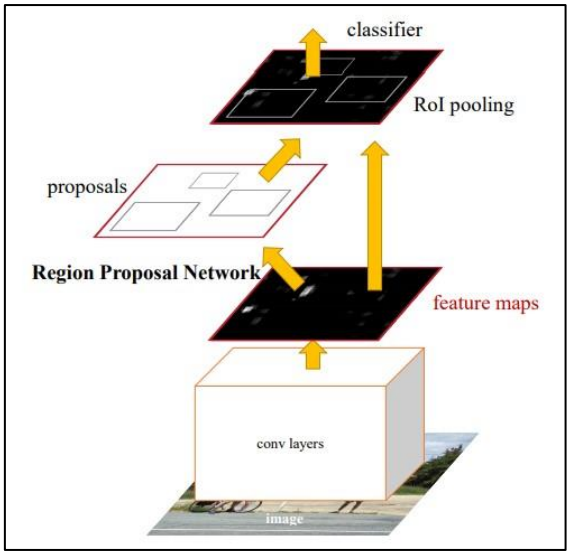

图 6 Faster R-CNN 结构[21]

四、时尚检索

时尚检索可分为几大类别，包括最流行的从图像或视频中寻找相同或相似的物品，以及时尚推荐的研究。目前只有些许处理物品属性的图像检索的研究，比如从图像和文本中发掘空间感知概念，并用 word2vec 模型进行属性反馈的物品检索[22]。

五、识别服装属性

服装属性是一种评估服饰的有效工具，一些基础工作，如 SIFT 和 HOG，与支持向量机结合以处理手工特征。另一种混合匹配（mix and match）的方法[23]将深度神经网络和条件随机场组合，来寻找服装产品和属性之间的兼容性。还有一些工作利用弱标记（weakly labeled）的图像-文本对来发现属性[24]。

第二章　模型

第一节 概要

现有的大多数研究会在一个全局空间中表示数据对象，都会用一个多维隐向量来表示，隐向量的每个元素的含义是未知的，正因如此，深度学习算法的可解释性和结果的说服力会稍显不足。在图 7 中，假设一个目标对象被分解到三个维度 x、y、z，可用(x, y, z)来表示，然而每个维度具体代表了什么是不确定的。针对这一点，需要一种新的空间，这个空间不再以隐维度为基准，而是由具体的物品属性组成，这些属性有真实的语义信息。在这里，物品实际上都是各类服饰，投影空间的各维度对应服饰的不同特征，例如一件衬衫，维度可能是袖长、领口、纽扣设计等等，每个维度都有不同的值，比如袖长可以是长袖或者短袖。通过这一过程，服饰被投影到各个维度上，服饰的语义属性也能在这个空间中表示出来，这样就可以获得视觉可解释的物品表征。

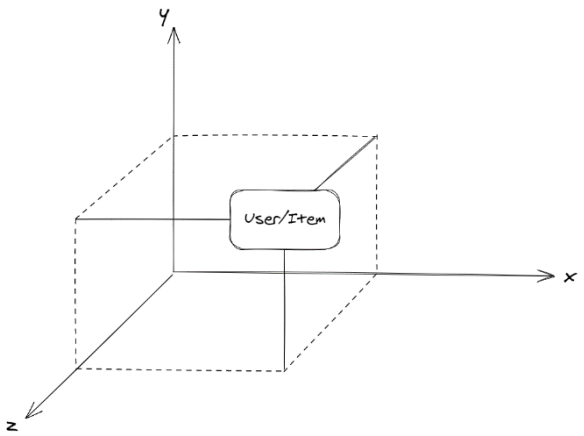

图 7 隐向量示意图

在这个模型中，分类能力（多任务分类网络）和特征的定位（Grad-CAM）使得确定属性的具体位置和类别变得可能。可以看到，模型通过生成 bounding box 彰显用户可能感兴趣的部分，再显示该区域属于哪个语义属性。有了这些结果，便可通过对其他除物品之外的特征进一步处理，比如计算用户对各种属性的偏好、用户对一种属性感兴趣的可能性等，将表征的物品转化为实际的推荐结果。

第二节 属性投影空间

这一节将描述如何把服饰物品投影到上述空间中，并获得物品属性的表示。如上所述，首先需要把服饰分为几个语义属性，不同的属性对应不同的服饰特征，再为这些特征生成表示。现实世界中，服装的各个特点只能用人眼观察，网络电商中服装贩卖的展示图并不包含服饰的详细信息，因此，需要用一个经过人工标

注的数据集来训练模型，数据集将在后文介绍。这个模型将对服饰图片进行分类，得到各标签的信息，然后定位代表这些标签的特征，最后生成服饰的特征表示。由于数据集中同样缺失 bounding box 的信息，因此不能像目标检测那样直接提取特征并获得表示。模型采用了前文 Grad-CAM 的方式来定位特征，即通过特征图上的梯度信息大致标出图片的特征，得到该特征的位置。

第三节 模型结构

该模型的基本结构示意如图 8 所示，先训练分类卷积网络得到 Grad-CAM，再通过后续的 ROI 操作获得服饰特征表示。该模型能够识别相关的特征，并且忽略那些相关性小的特征。

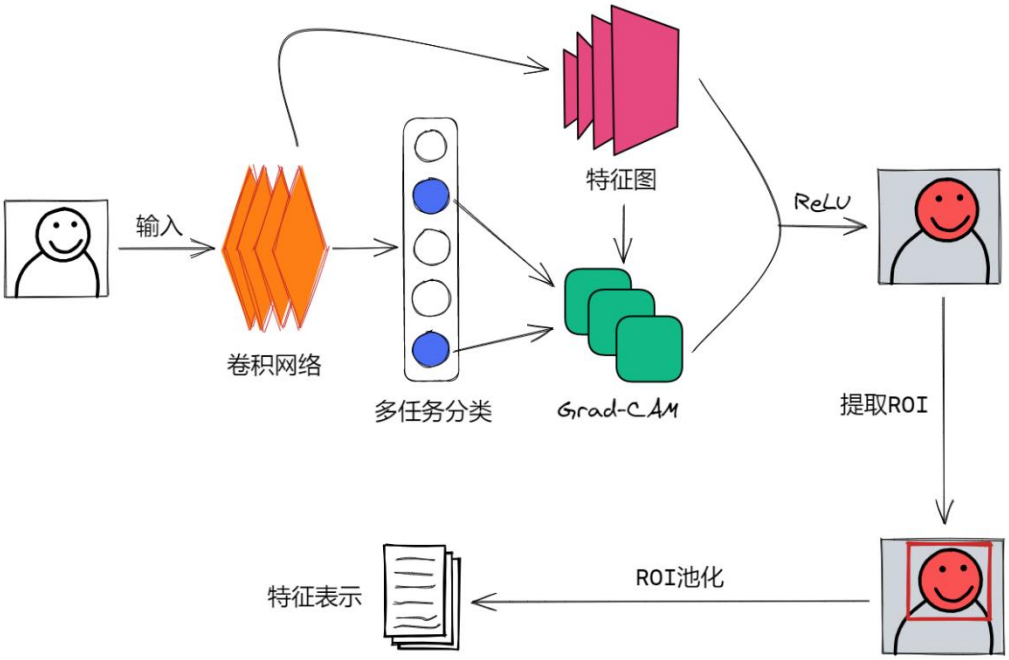

图 8 模型结构

一、卷积网络

模型中采用的卷积神经网络是 ResNet18[25]。根据之前的研究成果，深度神经网络随着层数的增加，网络可以提取到越来越复杂的特征，从直觉上说，网络越深，模型的效果也就越好。然而，事实却与直觉相反，网络的不断加深并未产生更好的结果，训练误差和测试误差反而都在增加，这可能是因为网络过深导致

的梯度消失或者梯度爆炸，神经网络不能正常收敛，从而发生了退化。

ResNet 设计的初衷便是为了解决退化问题，Res 指的是残差，一个基本的残差单元如图 9 所示。残差的核心思想是构造某种恒等映射，假设网络的输入为 x，期望输出为 H(x)，现在令

$$H(x) = F(x) + x,$$

对于网络而言，仅需学习残差

$$F(x) = H(x) - x。$$

网络学习残差 F(x)的效果要比直接学习原始特征 H(x)好很多。

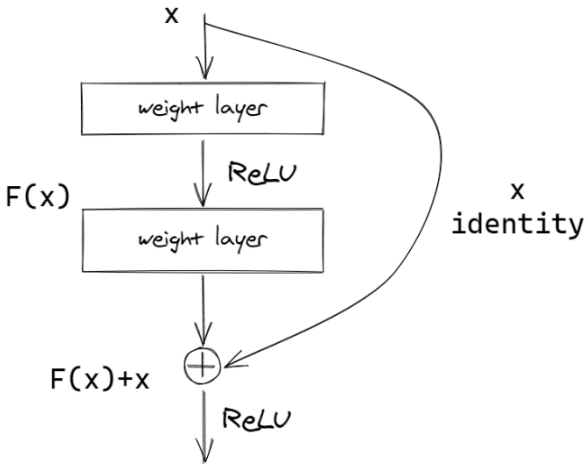

图 9 残差单元

ResNet 有多种结构，常用的有 ResNet18、ResNet34、ResNet50、ResNet101、ResNet152，这些结构在整体上大同小异，网络结构总览可以参照图 10，输入都由一个卷积层和一个最大池化层组成，不同结构的具体差异在于中间卷积的部分。对于 ResNet18，中间卷积共有四个模块，每个模块由两个残差块和一个减半采样构成，每个残差块又包括两个 3x3 卷积层，具体细节可参照图 11，可以看到在中间部分的四个模块中，残差块都重复堆叠了两次。网络输出部分即是通过平均池化将所有特征图转为 1x1 的大小，针对该模型，卷积网络最后输出的全连接层会分别对每个服装属性做设置以实现多标签的分类能力。

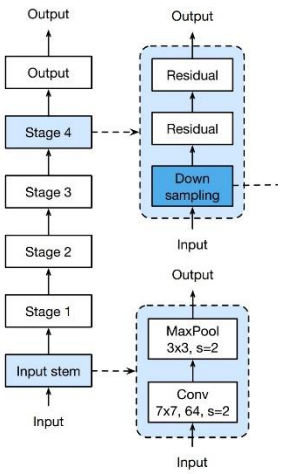

图 10 ResNet 结构

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

图 11 ResNet 网络参数，来源[25]

二、网络损失函数

一般情况下，分类任务有多种损失函数可供选择，常常会使用交叉熵来计算。熵是服从某一特定概率分布事件的理论最小平均编码长度。对于两种概率分布 P 和 Q，P 为真实概率分布，Q 为估计概率分布，首先熵的计算如下：

$$Ent(P) = \mathbb{E}_{x\sim P}[-\log P(x)]$$
$$Ent(Q) = \mathbb{E}_{x\sim Q}[-\log Q(x)]$$

如果 Q 即为真实概率，那么二者的熵显然相同，但此时 Q 并不一定为真实的概率分布，且$-\log P(x)$和$-\log Q(x)$最小编码长度也不同，这两个熵的对比可能毫无意义。因此，我们可用真实概率分布 P 计算平均编码长度，用估计概率分布 Q 计算实际编码长度，这就是交叉熵。交叉熵使得实际编码长度和理论最小平均编码长度的比较有了意义。交叉熵的计算：

$$CrossEntropy(P,Q) = \mathbb{E}_{x\sim P}[-\log Q(x)]$$

在该模型中，训练数据先通过一个卷积网络，通过多标签分类得到初步结果，因为一种属性可能同时属于不同的类，所以并未使用 softmax 层，卷积神经网络的损失采用 sigmoid 处理输出的交叉熵，即

$$loss = CrossEntropy(Sigmoid(output), labels))$$

通过最小化损失，我们可以找到服饰最大概率的几个标签，从而找到对应特征所在的区域。图 8 中多任务分类下蓝色部分示意为分类结果。

三、含有梯度信息的类别激活图

在此之后，利用训练好的分类网络，我们便可以对每个特征计算它的 Grad-CAM（利用梯度信息产生的类别激活图）。首先，我们已经得到了卷积网络产生

的特征图，这里我们需要最后一个卷积层的特征图。按照 Grad-CAM 的原理，需要根据每个特征图的梯度信息计算出对应的权值，最后将所有特征图加权求和即可得到结果。在 Grad-CAM 的原论文中，权重用如下方式计算：

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

其中，y^c表示类别c对应的梯度，A^k表示第k张特征图（即第k个通道），求偏导相当于梯度反向传播，求和过程相当于全局平均池化，(i,j)事实上表明了最后一个卷积层中该特征图的位置。权值α_k^c表示了一种部分线性化（partial linearization），代表特征图对于目标类别的贡献，也就是重要性。

接下来便可求得 Grad-CAM：

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k)$$

这是一个和卷积网络特征图同样大小的热图，每个元素说明了类别c对该激活图的贡献。由于我们只关注那些对分类有正影响的特征，因此在线性化后采取了 ReLU 的操作，负影响的点可能代表了其他类别。在图 8 中 ReLU 步骤之后的结果里，红色部分表示贡献度最大的部分。

四、感兴趣区域的提取和池化

之后，利用得到的 Grad-CAM 图象，我们找到模型已经标注的部分作为 bounding box，这部分就是图象中最大的连通域。ROI（Region of Interest，感兴趣区域）部分将特征图上的重要部分汇集到一起，有了 bounding box 和最后一个卷积层的特征图，我们就可以通过 ROI 池化获得和区域相关的服饰特征表示。ROI 池化通过最大池化将 ROI 中的特征转换到一个较小尺寸的特征图中，这个特征图的尺寸是固定的，它的宽高独立于任何 ROI，即是说特征图的大小是超参数。实现过程中，所

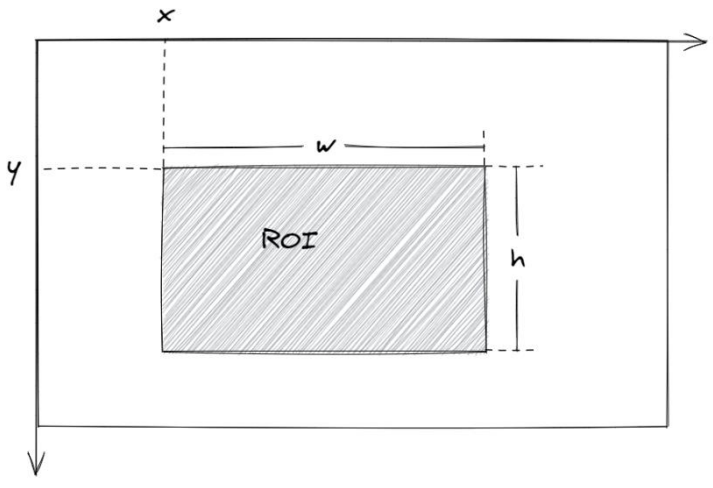

图 12 ROI 表示

有的 ROI 都是一个矩形区域，并且用坐标[x, y, w, h]表示，如图 12 所示。这种最大池化把原始的$h \times w$ ROI 区域划分为$H \times W$个子窗口，每个子窗口的大小为$h/H \times w/W$，这些子窗口中的最大值汇集到对应的输出窗口。ROI 池化和标准的最大池化相同，对于每个输入图像的通道都是独立的。图 8 里提取对应的 ROI 后生成了一个 bounding box（红色框），池化后便得到属性表示。

第三章　实验

这一章将展示模型各个部分的实验结果，以及一些对模型的考虑。

第一节 数据集

本次实验采用的数据集来自阿里巴巴的天池大数据竞赛[1]，数据集中每个可识别的服装标签都需要检测，考虑到时尚相关知识的复杂性，数据集中只包含单个主体的产品图像。

一、标签

数据集中所有图像数据都经过了人工标注。为了保持每个属性维度的均匀性，有些标签信息是故意缺失的，比如一件衣服既可观察到衣领设计同时也能看到袖子长度，那么只会标注衣领或者袖长的信息，而不会同时记录两个属性的标签。图像标签选择了 8 种属性，在表 1 中说明。

表 1 图像标签

属性	类别
coat length	invisible, high waist length, regular length, long length…
collar design	invisible, shirt collar, Peter Pan, Puritan collar…
lapel design	invisible, notched, collarless, Shawl collar…
neck design	invisible, turtle neck, ruffle semi-high collar, low turtle neck…
neckline design	invisible, strapless neck, deep V neckline, straight neck…
pant length	invisible, short pant, mid length, 3/4 length, cropped pant…
skirt length	invisible, short length, knee length, midi length, ankle length…
sleeve length	invisible, sleeveless, cup sleeves, short sleeves, elbow sleeves…

二、数据特点

服装图片在特定属性维度下的属性值都是互斥的，不会有相互冲突的属性值，例如在高领设计中，标签不能同时为高领和荷叶边半高领。为了保证属性值互斥，所有图片中不会出现重叠的服装。

[1] https://tianchi.aliyun.com/competition/entrance/231671/information

不同属性维度下的属性值可以在同一张图片中共存，但不会同时标注，比如 shirt collar 和 turtle neck。

在表 1 中可以看到，每个属性维度下，都有一个“invisible”值。这表明一个特定的属性实际存在，但在具体的图像中没有出现或者被遮挡。例如，裙子的下摆被遮住了，所以裙子的长度尺寸将被标记为 “invisible”。

三、数据标注

一张图片的标注信息可以参考表 2。表 2 中 ，属性值有 9 个字符，分别对应了袖长的 9 种类型（包括“invisible”），其中“y”表示“肯定是”，“n”表示“肯定不是”，“m”表示“可能是”，一张图片有且仅有一个“y”，其他属性值可以为“n”或“m”。之所以有这三种情况，是因为现实世界中服装类别模糊不清、无法完全确定是非常正常的，不可能有一个严格的标准来判断一件衣服属于长袖还是超长袖，它可能比其他长袖要长，但又比其他超长袖要短，如果这件衣服相比于超长袖更接近长袖，那么它会被标注为长袖“y”、超长袖“m”。

实验过程中，考虑到一个用户可能仅仅喜欢袖子较长的衣服，而不关心究竟是长袖还是超长袖，因此把所有标为“m”对应的标签以 0.5 的概率改为“n”或“y”，模型分类的阈值设为 0.6，即有 40%分为对应类别，剩余 60%不归于该类。

在有遮挡的情况下，“invisible”标签会被设为“y”，其他值都为“n”。

表 2 标注信息

图片名称	属性	属性值
0012345abcdef.jpg	sleeve_length_labels	nnnnnnnym

四、预处理

将“n”和“y”分别改为 0 和 1，属性值由字符串改为对应的 0-1 向量。在图像数据输入之前，因为绝大部分图片的尺寸为 512x512，仅少部分并非这个尺寸，所以全部图片的尺寸都缩放到 512x512，然后做了随机裁剪和旋转等增强操作，最后再将图片归一化。

第二节 分类网络

同样由于算力限制，网络仅训练了 10 轮，在正式训练之前进行了预训练，因此权重等参数并非从初始状态开始，一些指标也并非从 0 开始。学习率设为 0.01，权重衰减率设为 0.0005，模型采用随机梯度下降优化，使用精确率（precision）作为评价指标。

一、模型损失

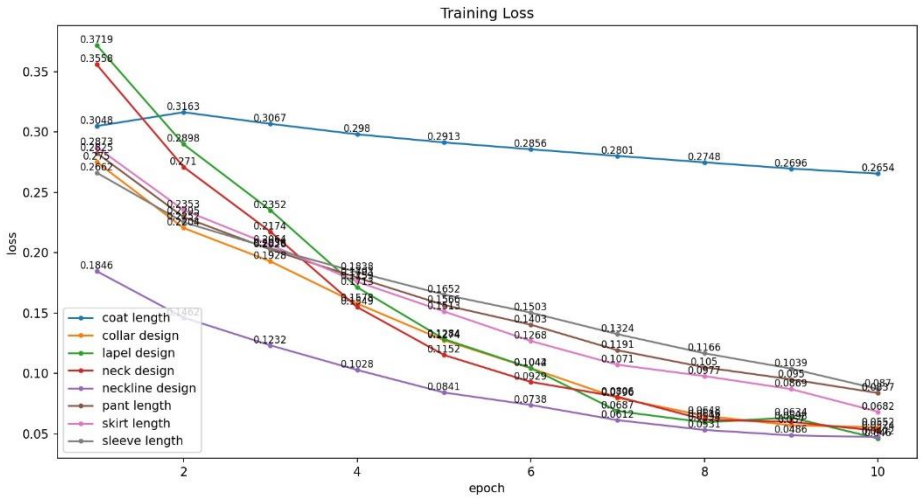

图 13 训练损失

表 3 测试损失

属性	Loss
coat length	0.2815
collar design	0.2929
lapel design	0.2337
neck design	0.2627
neckline design	0.1401
pant length	0.3177
skirt length	0.2440
sleeve length	0.2966

图 13 展示了模型在训练集上的损失，表 3 列出了模型在测试集上的损失。从图表中可以看出，大体上所有属性的分类损失都在下降，“coat length”属性的

训练效果不是很好，损失保持在 0.26~0.31 的范围之间，未有明显下降，其他属性的优化效果都较为显著。再看测试集，所有属性相比于训练集的损失都不低，可能模型存在一定程度的过拟合现象。

值得注意的是，从图中看来，“design”类型的属性损失最后都能降至较低的值，而“length”类型却相对高一些。猜想是因为长度相关的属性比较模糊，比如裤子长度是应该考虑从腰部到裤脚的整体还是只需考虑裤脚的位置，如此一来，模型对“length”属性可能会产生两种判别方式，导致分类上的不一致。

二、分类精确率

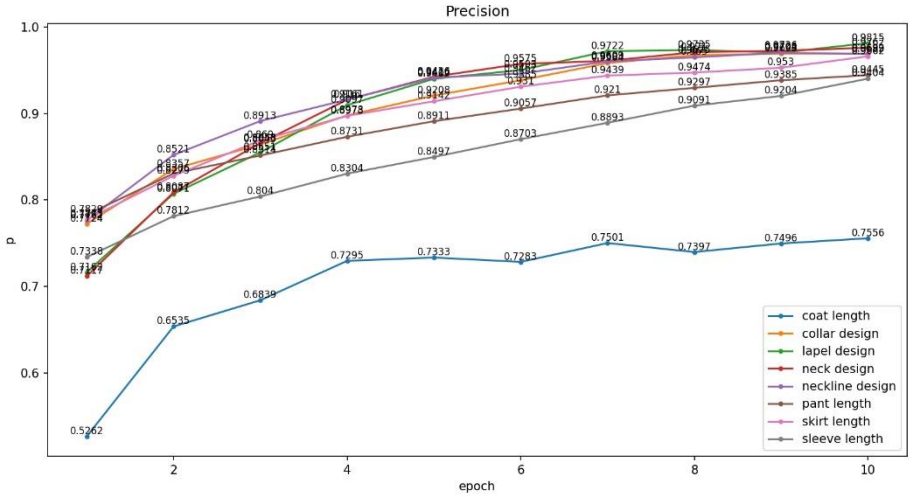

图 14 精确率

表 4 测试集精确率

属性	Precision
coat length	0.7009
collar design	0.8166
lapel design	0.8790
neck design	0.8240
neckline design	0.8329
pant length	0.7456
skirt length	0.8161
sleeve length	0.5847

从精确率来看，“coat length”属性的分类表现依然不佳，其他属性在训练时的精确率最终都能达到 0.95 左右。在测试集上，“sleeve length”和“coat length”属性的精确率都不高，其他属性基本能达到 0.8 以上。依然能看到，“length”类属性的表现比“design”类还是要差一点，所以长度判断的基准是非常有可能对分类器性能产生影响的。

第三节 特征可解释性

这一节将展示从服装图像生成激活图的样例。先来看一下“collar design”图像的生成，在图 15 中，左图为原始图像，右图是生成的激活图。

图 15 Grad-CAM 效果图

对于“collar design”属性，显然模型应该关注的是衣领，在这张图片中，模特所穿衣服的领口处正是模型的标记区域，也就是对分类结果贡献度最大、包含信息最多的区域。这张图对应的特征热图可以参考图 16。不同属性的激活图效果大致类似，基本上最重要的区域都由模型识别。

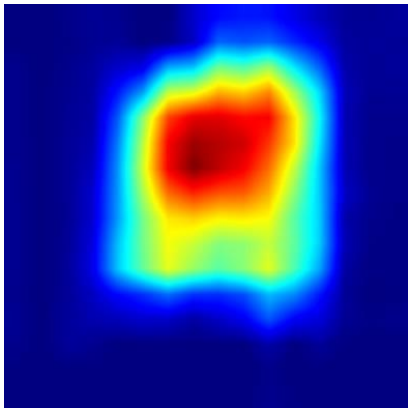

图 16 热图效果

模型的这一阶段会将上一步中类别激活图关键的区域作为 ROI 提取出来，以图 12 中的方式保存 bounding box 的位置信息，再以“neckline design”属性为例，效果比较如图 17 所示，左图依然为原图，右图为提取到的 bounding box。如前所述，bounding box 是一个矩形区域，有图有所拉伸是因为在模型运行过程中输入的原始图片尺寸全部调整为 512x512，绝大多数图片大小无需改变，有些图片（例如该样图）并非正方形，因此图片比例会与原始图片不一致。

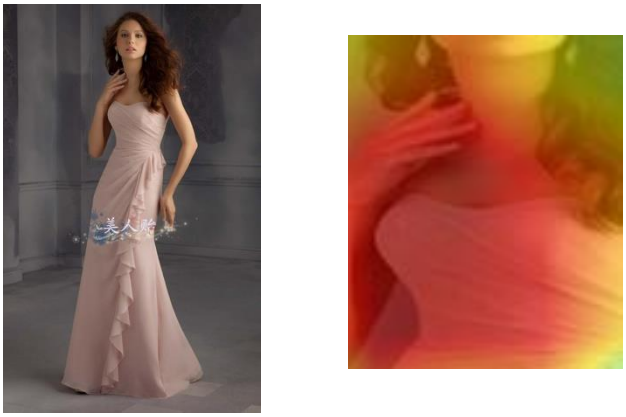

图 17 ROI 提取效果

第四节 特征表示

经过 ROI 池化后，便可得到 bounding box 部分的特征表示，可以设置任意的向量形状，这里以（7, 7）大小的特征向量为例，大致如图 18 所示，每张图片的特征转化为浮点形式，图中仅为部分结果。

图 18 特征表示向量

实际上，每个特征表示向量的形状为（1,3,7,7），“1”代表这一张图片，“3”代表三个颜色通道 R、G、B，7x7 则是每个图片中的特征表示，在应用到其他模型中时，可根据实际要求改变特征向量的大小。

第五节 对图像可解释的思考

图像的可视化中有一种权衡，一个有良好可解释性的可视化图像并不能准确代表模型，反之，能够体现模型可靠性的图像却并不容易可视化，或者可视化后有清晰的可解释性。可靠性来自对模型的全部描述，而这却是可视化的难点，即使成功可视化，也可能是人们不能理解的图像。图 19 中从左到右分别是原图、激活图、热图，图 20 中是对应的裁剪后的版本，属性为“pant length”。

图 19

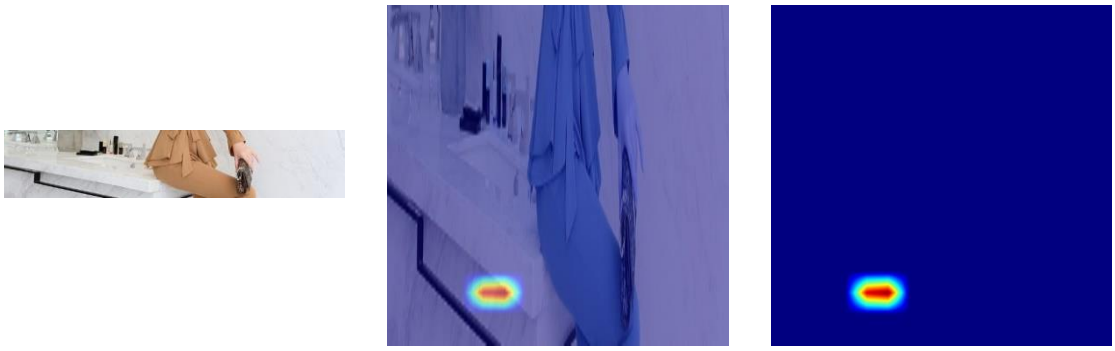

图 20

在关键区域被“遮挡”后，模型的识别效果显然变得很差，根据相同的损失计算方式所得的结果为 0.7131，几乎是测试集损失的两倍，比训练集上的损失大了一个数量级。在推荐系统中，这当然是无法捕捉到用户所偏好的服装特点并产生有效推荐结果的。对于深度模型来说，解释模型或者可视化并不是那么容易。

从以上的结果来看，可以认为这种方法是合理的，也是可解释的。通过原始图像和裁剪图的比较，可以看到 Grad-CAM 可以较好地解释模型，有着良好的可靠性。

第六节 总结

在这个模型中，我们可以通过图像分析了解各种物品的细粒度特征，并用物品的属性表示为基于用户偏好的推荐打下基础，同时也保证了物品推荐的良好可解释性。为了达成这一目的，该模型将物品投射到一个可解释空间中，这个空间的每个维度都代表了现实世界里真实的语义属性，并在其中学习物品的特性表征，模型在数据集上的表现证明了其可解释性。

通过 Grad-CAM 生成的视觉说明，这个基于卷积网络的模型对人来说变得更加透明。Grad-CAM 的定位能力使其能够识别与对应属性最相关的区域，对于不同属性，可以利用对应的网络结构适应。这种方法的可视化不仅拥有良好的可解释性，在可靠性方面也有保证。经过实验，Grad-CAM 可以有效的区别不同类别的属性，将其利用在目标检测相关的技术中同样能提供清晰的解释来说明模型的输出结果，为模型提供了可靠的解释。

如果有未来的研究，将该模型应用到时尚推荐之外的其他领域也是一个有趣的课题。真正的人工智能不仅应该是智能的，还需要让人能够理解它的行动和信念，从而让人能充分相信它的决策和决定，为了达成这一目的，模型的可解释性是必不可缺的。

参 考 文 献

[1] J Matas, O Chum, M Urban, T Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and Vision Computing*, Volume 22, Issue 10, 2004, Pages 761-767, ISSN 0262-8856.

[2] Susanta Sarangi, Md Sahidullah, Goutam Saha, “Optimization of data-driven filterbank for automatic speaker verification,” *Digital Signal Processing*, Volume 104, 2020, 102795, ISSN 1051-2004.

[3] Shigeki, Sugiyama (12 April 2019). “Human Behavior and Another Kind in Consciousness: Emerging Research and Opportunities.” IGI Global. ISBN 978-1-5225-8218-2.

[4] Rolnick, David & Tegmark, Max. (2018). “The power of deeper networks for expressing natural functions.” *International Conference on Learning Representations*. *ICLR 2018*.

[5] Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.-Z. (2019-12-18). “XAI-Explainable artificial intelligence”. *Science Robotics*. 4 (37): eaay7120.

[6] O. Loyola-González, “Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View,” in *IEEE Access*, vol. 7, pp. 154096-154113, 2019.

[7] R. Roscher, B. Bohn, M. F. Duarte and J. Garcke, “Explainable Machine Learning for Scientific Insights and Discoveries,” in *IEEE Access*, vol. 8, pp. 42200-42216, 2020.

[8] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 3 (May-June 2018), 31–57.

[9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, Volume 58, 2020, Pages 82-115, ISSN 1566-2535.

[10] Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, “Methods for interpreting and understanding deep neural networks,” *Digital Signal Processing*, Volume 73, 2018, Pages 1-15, ISSN 1051-2004,

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, “Learning Deep Features for Discriminative Localization,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921-2929*.

[12] K. E. Ak, A. A. Kassim, J. H. Lim and J. Y. Tham, “Learning Attribute Representations with Localization for Flexible Fashion Search,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7708-7717*.

[13] F. Xiao and Y. J. Lee, “Discovering the Spatial Extent of Relative Attributes,” *2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1458-1466*.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626*.

[15] Julian McAuley and Jure Leskovec. 2013. “Hidden factors and hidden topics: understanding rating dimensions with review text.” *In Proceedings of the 7th ACM conference on Recommender systems (RecSys '13)*. Association for Computing Machinery, New York, NY, USA, 165–172.

[16] Gao, Jingyue & Wang, Xiting & Wang, Yasha & Xie, Xing. (2019). “Explainable Recommendation through Attentive Multi-View Learning.” *Proceedings of the AAAI Conference on Artificial Intelligence. 33. 3622-3629. 10.1609/aaai.v33i01.33013622*.

[17] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. “Image-Based Recommendations on Styles and Substitutes.” *In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 43–52.

[18] Qiang Liu, Shu Wu, and Liang Wang. 2017. “DeepStyle: Learning User Preferences for Visual Recommendation.” *In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 841–844.

[19] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. “Aesthetic-based Clothing Recommendation.” *In Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 649–658.

[20] R. Girshick, “Fast R-CNN,” *2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440-1448*.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. “Faster R-CNN: towards real-time object detection with region proposal networks.” *In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15)*. MIT Press, Cambridge, MA, USA, 91–99.

[22] X. Han et al., “Automatic Spatially-Aware Fashion Concept Discovery,” *2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1472-1480*.

[23] Kota Yamaguchi, Takayuki Okatani, Kyoko Sudo, Kazuhiko Murasaki and Yukinobu Taniguchi. “Mix and Match: Joint Model for Clothing and Attribute Recognition.” In Xianghua Xie, Mark W. Jones, and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC), pages 51.1-51.12*. BMVA Press, September 2015.

[24] S. Vittayakorn, T. Umeda, K. Murasaki, K. Sudo, T. Okatani, and K. Yamaguchi. “Automatic Attribute Discovery with Neural Activations.” *European Conference on Computer Vision (ECCV), 2016*.

[25] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.