计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800209

基于语言描述的细粒度美妆图片排序

姚林丽　陈师哲　金　琴
中国人民大学信息学院　北京 100872
(linliyao@ruc.edu.cn)

摘　要　文中研究了化妆领域中基于文本的细粒度视觉推理问题,具体探究了一个新颖的多模态任务,即根据有序的化妆步骤描述,对化妆过程中打乱顺序的人脸图片进行排序。针对这个新颖的任务,通过数据的处理和分析,提出了两个排序模型:第一个排序模型从单模态的角度出发,只利用图片的信息进行排序;第二个模型从多模态的角度出发,通过建立文本描述和图片之间的联系来指导图片排序。在 YouMakeup VQAChallenge 数据集上进行了详实的实验以及分析,实验结果表明,所提出的两个模型在不同的图片对数据上具有互补性,在美妆图片排序任务上具有良好的表现,在测试集上的选择准确率分别达到了70%和58.93%。

关键词:图片排序;视觉推理;美妆领域;细粒度;多模态;深度学习
中图法分类号　TP37

Fine-grained Facial Makeup Image Ordering via Language

YAO Lin-li,CHEN Shi-zhe and JIN Qin
School of Information,Renmin University of China,Beijing 100872,China

Abstract　This paper studies text-based fine-grained visual reasoning in makeup domain and explores a novel multi-modal task, which sorts a set of facial images from a makeup video into the correct order according to the given ordered step descriptions. On this novel task,this paper first does data processing and analysis to learn the characteristic of the makeup dataset,and then proposes two baseline models to solve the image ordering task. The first baseline model only uses image information and ignores the guiding role of the text description from a single-modal aspect. The second model utilizes the text semantics to guide image ordering,establishes the relationship between text description and images and can reason the visual appearance change brought by step description. This paper conducts extensive experiments on the YouMakeup VQA dataset. The experiments show that the two models are complementary to each other,and achieve good performance on the image ordering task,with the selection accuracy on the test set of 70% and 58.93% respectively.

Keywords　Image ordering,Visual reasoning,Makeup domain,Fine-grained,Multi-modal,Deep learning

1　引言

随着互联网上图片、文字等多模态数据的激增,以及图像和自然语言处理技术的突破,融合图像和文本的多模态任务已然成为研究热点,例如图像字幕生成、视觉问答任务、基于语言的视觉导航等。其中,理解语言描述对视觉物体的影响是一项重要且应用广泛的研究,例如用户可直接利用自然语言与机器交互来修改物体的视觉状态。但是,由于缺乏数据集和有效的建模方法,这一问题并未得到充分的研究。因此,本文研究了一个新颖的多模态任务,在化妆领域中,基于文本进行细粒度的视觉推理。这个任务由 YouMakeup VQA Challenge[1] 提出,定义为美妆图片排序,即根据有序的化妆动作描述,对化妆过程中打乱顺序的人脸图片进行排序。具体地,该任务的形式是视觉问答,每个问题包括 5 张人脸图片和数目不定的英文化妆步骤描述,4 个候选答案表示不同的图片顺序,任务目标是选出正确的答案,即对图片进行排序,如图 1 所示。此任务需要模型捕捉文本动作描述和视觉图片之间的相关性,进而理解不同化妆动作带来的视觉变化,最终推理出图片的正确顺序。

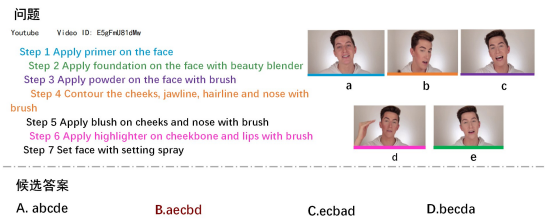


图 1　细粒度美妆图片排序任务示例图

Fig. 1　Example of fine-grained facial image ordering task

到稿日期:2020-07-30　返修日期:2020-09-06　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61772535);北京市自然科学基金(4192028);国家重点研发计划(2016YFB1001202)
This work was supported by the National Natural Science Foundation of China(61772535),Natural Science Foundation of Beijing(4192028) and National Key Research and Development Plan(2016YFB1001202).
通信作者:金琴(qjin@ruc.edu.cn)

在美妆这一特定领域，基于文本的细粒度视觉推理任务具有充分的研究和应用价值。在研究价值方面，化妆过程天然具有细粒度的特性：不同的化妆步骤都作用在同一张人脸上，因此它们拥有相似的背景，但同时每个化妆步骤之间都有明显或细微的区别，适合于挖掘化妆动作带来的细粒度人脸视觉变化。在应用价值方面，近年来社交媒体和短视频平台快速发展，促进了美妆产品的宣传，大力带动了美妆产业的发展，而迅速壮大的美妆产业和巨大的消费市场也反过来催生了美妆产品的搜索、推荐、分享等衍生功能。在此背景下，研究不同化妆步骤带来的人脸视觉变化，可以促进机器对化妆过程的理解，具有广泛的应用价值，例如辅助美妆博主进行美妆图片的编辑，为用户推荐合适的妆容教程等。

化妆是一种特殊的艺术形式，它作用于“人脸”这张特殊的画布。它借助口红、眼影、粉底等工具，依靠化妆师的直觉和经验，来定制面部几何形状、肤色、局部色彩，凸显独特的个人特征，以突出和提高一个人的自然美[2]。在化妆领域，现有的工作主要聚焦于为人脸增添妆容或删除妆容这个双向过程[3]，例如化妆风格迁移[4-5]为素颜的人脸添加指定风格的妆容，而在鲁棒的人脸识别应用[6]中需要去除面部妆容。这些工作的关注点在于妆容的“有无”，而本文的工作更细粒度地分解了化妆过程，致力于理解每个化妆步骤带来的妆容变化。虽然熟悉面部固有形态的观察者，通过观察完整妆容的效果，容易得出主观的评价[2]，但是观察每一个化妆步骤带来的视觉差异是有难度的。与此同时，文化和社会趋势的变迁引起化妆风格不断地演变，多样的风格也为排序化妆过程中的图像增加了不确定性。

本文在这个新颖的任务上提出了两个模型，并进行了详实的实验分析。具体地，第一个模型从单模态的角度出发，只利用了图片的信息而忽略了文本描述的指导作用。首先，在化妆步骤未知的情况下，探究模型仅利用人脸视觉信息预判图片顺序的表现，并将其作为一个先验参照。然后，将核心问题建模为一个图片分类任务，输出类别对应两张图片的相对顺序。第二个模型从多模态的角度出发，通过建立文本描述和图片之间的联系来指导图片排序。该模型将核心问题转换成一个组合式图片检索任务，即用一张初始人脸图片加一句化妆步骤描述，来检索或者匹配化妆步骤完成后的人脸图片。通过提升检索的精度来促使模型学习步骤描述带来的相关视觉变化。实验结果表明，所提出的两个模型在处理间隔不同化妆步骤的图片对数据时，具有互补性；在美妆图片排序任务上具有良好的表现，在测试集上最高可以达到 70% 的选择准确率。

2 相关工作

2.1 YouMakeup VQA Challenge

YouMakeup VQA Challenge[1] 是 CVPR 2020 承办的一个 workshop 比赛，旨在对美妆领域的视频进行细粒度的动作理解。比赛基于一个大规模、多模态、细粒度的美妆视频数据集 YouMakeup[7] 进行。

比赛的一个子任务是细粒度美妆图片排序，问题定义为：根据有序的化妆步骤描述，对来自同一个美妆视频、不同步骤的人脸图片进行排序。任务以视觉问答的形式给出：给定 5 张人脸图片和数目不定的英文化妆步骤描述，参赛者需要在官方提供的 4 个候选答案中选出正确的图片顺序。任务以选择准确率为模型表现的评判指标。若要解决这个美妆图片排序任务，需要细粒度地理解不同化妆动作带来的人脸形态变化。

2.2 组合式图片检索

Vo 等[8]提出了一种新的图片检索任务，即用一张原图片加一个包含修改信息的句子来检索一张符合条件的目标图片。这个图片检索任务属于跨模态检索的范畴，与被广泛研究的文本-图片检索任务[9-13]有所区别，其核心是实现图片-文本组合与目标图片之间的匹配。这种检索模式在实际生活中有丰富的应用场景，例如电子商务的产品搜索可以将用户的反馈整合到查询项中以提升图片检索的质量。Vo 等[8]提出了一种残差门控模块（Text Image Residual Gating，TIRG）来解决这个组合式图片检索任务，实现了在特征向量空间中用文本特征修改源图片特征，使源图片特征与目标图片特征的距离更接近。Guo 等[14]研究了一个相似的任务，即在电子购物的场景下，智能客服需要根据原商品图片和顾客给出的文字反馈返回一张符合条件的新商品图片。在这个任务上，他们更关注人和机器之间的多轮交互性。Mehrdad 等[15]从更细粒度的角度出发，显式地建立了图片区域和文本单词之间的关系，提升了模型的表现。此外，Park 等[16]和 Tan 等[17]提出了一个逆向任务，通过生成文本来描述两张图片之间的差异。

2.3 课程学习训练策略

课程学习（Curriculum Learning，CL）是一种用于机器学习的训练策略，最早由 Bengio 等[18]提出。在学习过程中，当学习的样本不是随机出现，而是按照一个有意义的顺序组织起来时，人和动物能够学得更好。例如，人类的成长需要经过很多年的在校学习，而教育系统是高度体系化和组织化的，让人从易到难、循序渐进地获得知识。受到人和动物的启发，Bengio 等将这种由易到难、循序渐进的学习策略应用于机器学习中，以提升模型的表现。在机器学习中，这种课程学习的策略具体体现为让模型先从简单的概念、数据或子任务开始学习，然后逐渐增强训练数据或者任务的难度，通过多阶段的课程训练来加快模型的收敛，并提升模型的泛化能力。

3 模型介绍

3.1 纯图片排序

化妆是一个不断往脸上的各个部位“增加”化妆品的正向过程，即使化妆步骤是未知的，我们也能利用图片中人脸形态的信息大致判断出图片的顺序。例如，观察到图片 a 中唇部呈淡粉色，图片 b 中唇部呈鲜红色，则可以推断出图片 b 在图片 a 的基础上叠加了“口红”，因此图片 b 的顺序应排在图片 a 之后。为了探究这种化妆先验知识能达到的排序效果，首先单纯利用图片的视觉形态信息，来解决图片排序任务。

本文将核心任务转换为两张图片之间顺序的比较，提出了图片对比较模型（记为 Pairwise），其本质是一个图片二分类器。训练好的图片对比较模型可以预测两张图片的相对顺序，我们再将其应用于美妆图片排序任务。

3.1.1 图片对比较模型的结构

基于孪生网络的图片对比较模型的结构如图 2 所示。该模型主要包含一个孪生网络[19]和一个二分类器。其中，孪生网络由两个图片特征编码器(ResNet-18)构成，两者具有相同的结构，且在训练中共享参数。二分类器包含两层全连接层和 Sigmoid 激活函数，最终输出类别为 0 或 1。输入两张来自不同化妆步骤的图片(I_i, I_j)，我们用孪生网络提取两张图片的特征向量，随后拼接这两个特征向量并将其作为二分类器的输入。如果图片I_i在化妆过程中出现的时间更早，二分类器的输出类别为 1；反之图片I_j出现的时间更早，二分类器的输出类别为 0。我们用交叉熵损失函数来训练二分类器。

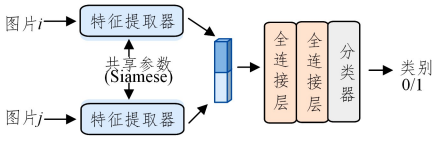


图 2 基于孪生网络的图片对比较模型

Fig. 2 Overview of image pairwise comparison model

3.1.2 选择题策略

本文使用训练好的图片对比较模型来完成美妆图片排序任务。已知图片对比较模型可以预测两张图片的相对顺序，输出类别为 0 或 1，并得到每个类别对应的置信度。首先，针对每个候选答案，构造所有可能的有序图片对组合。由于每个候选答案表示 5 张图片的顺序，因此一共可以得到C_5^2个有序图片对组合。例如，在某个候选答案中，5 张图片的顺序依次为$\{I_1<I_2<I_3<I_4<I_5\}$，那么可以构造得到C_5^2个有序图片对$\{(I_1<I_2),(I_1<I_3),(I_1<I_4),(I_1<I_5),(I_2<I_3),(I_2<I_4),(I_2<I_5),(I_3<I_4),(I_3<I_5),(I_4<I_5)\}$。然后用训练好的模型预测每个有序图片对出现的概率，对C_5^2种组合累加求平均得到一个最后的概率分数，挑选出概率分数最大的候选答案作为最终答案。

3.2 基于文本的图片排序

单纯利用图片信息来推理化妆中人脸形态的变化是不够精准的，而化妆步骤描述可以在这个过程中提供额外的指导信息。例如，仅从宏观角度观察整张人脸，很难发现画眼线这类细微化妆动作带来的视觉变化；此时如果添加步骤描述“在睫毛根部填充眼线”能指导我们更精准地捕捉到眼睛的变化。因此，本文建立文本和图片之间的联系，在步骤描述的指导下对图片进行排序。

我们将核心问题转换为一个组合式图片检索任务：查询项由一张源图片和一句文本描述组成，目标是检索一张新图片。在化妆场景下，这种检索模式对应为：用一张初始的人脸图片加一个化妆步骤描述，来检索人脸形态改变后的相关图片，如图 3 所示。我们为这个任务构造三元组(I_i, S_i, I_j)，其中I_i表示未进行化妆步骤S_i时的人脸图片，I_j表示完成化妆步骤S_i后的图片。若模型要根据步骤S_i和源图片I_i准确地检索到目标图片I_j，则它首先要理解步骤S_i作用在图片I_i上带来的视觉变化，并能找到变化后的目标图片I_j。

首先在组合式图片检索任务上训练模型，使它能够理解步骤描述带来的相关视觉变化，随后将其应用于图片排序任务。

图 3 组合式图片检索示例

Fig. 3 Example of compositional image retrieval

3.2.1 组合式图片检索模型的结构

在组合式图片检索任务上，我们采用了文献[3]提出的经典模型 Text Image Residual Gating(TIRG)。该模型使用文本特征来“修改”原始图片特征，得到一个融合文本和图片信息的组合特征，训练目标是使得组合特征和目标图片特征在高维向量空间中距离更近。

TIRG 模型的整体结构如图 4 所示。该模型包括 3 部分：1)图片特征编码器；2)文本特征编码器；3)组合图片和文本特征的 TIRG 模块。首先，通过图片特征编码器得到图片特征，通过文本特征编码器得到文本特征；然后，通过 TIRG 模块“修改”源图片特征得到组合特征；最后，计算组合特征和待检索图片特征之间的匹配分数，以检索最相关的目标图片。

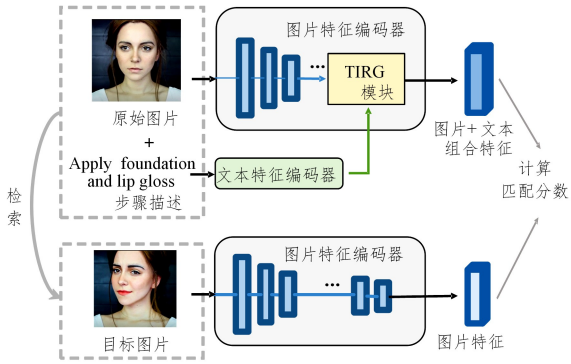


图 4 组合式图片检索模型 TIRG

Fig. 4 Overview of compositional imageretrieval model

(1)特征提取器

对于图片特征提取器，使用卷积神经网络 ResNet-18[20]提取图片的二维空间特征向量，用符号$\boldsymbol{\varnothing}_x$表示。它的维度是$(W, H, C)$，其中，$W$ 表示宽度，H 表示高度，C 表示通道数(共计 512 个通道)。对于文本特征提取器，采用循环神经网络 LSTM[21]编码输入的文本 t，取 LSTM 最后时刻的隐状态作为句子特征，用符号$\boldsymbol{\varnothing}_t$表示，它的维度是 512。

(2)组合特征模块

Text Image Residual Gating (TIRG) 模块使用残差-门控机制将图片特征$\boldsymbol{\varnothing}_x$和文本特征$\boldsymbol{\varnothing}_t$组合为一体，得到的组合特征用符号$\boldsymbol{\varnothing}_{xt}^{rg}$表示，计算方法如式(1)所示：

$$\boldsymbol{\varnothing}_{xt}^{rg}=w_g f_{\text{gate}}(\boldsymbol{\varnothing}_x,\boldsymbol{\varnothing}_t)+w_r f_{\text{res}}(\boldsymbol{\varnothing}_x,\boldsymbol{\varnothing}_t) \tag{1}$$

其中，f_{gate}表示门控特征，计算方法如式(2)所示；f_{res}表示残差特征，计算方法如式(4)所示；$\boldsymbol{w}_g$和$\boldsymbol{w}_r$是待学习的参数矩阵，用来平衡两种特征。

$$f_{\text{gate}}(\boldsymbol{\varnothing}_x,\boldsymbol{\varnothing}_t)=\sigma(f_{\text{layer}}(\boldsymbol{\varnothing}_x,\boldsymbol{\varnothing}_t))\cdot\boldsymbol{\varnothing}_x \tag{2}$$

$$f_{\text{layer}}(\boldsymbol{\varnothing}_x,\boldsymbol{\varnothing}_t)=W_{g2}*RELU(W_{g1}*[\boldsymbol{\varnothing}_x,\boldsymbol{\varnothing}_t]) \tag{3}$$

门控单元将输入的图片特征$\boldsymbol{\varnothing}_x$视为输出特征的一个主要参考，使得输入图片特征和输出组合特征存在于同一个特征向量空间。其中，σ 表示 Sigmoid 函数，$\cdot$ 表示元素积，$*$ 表示带批标准化(Batch normalization)的二维卷积操作，W_{g1}和

W_{g2}是3×3的卷积核，[·]表示拼接操作。由于图片特征$\boldsymbol{\varnothing}_x$和文本特征$\boldsymbol{\varnothing}_t$的维度是不同的，在实际操作中，先将$\boldsymbol{\varnothing}_t$的长和宽扩展成与$\boldsymbol{\varnothing}_x$相同的维度。

$$f_{\text{res}}(\boldsymbol{\varnothing}_x,\boldsymbol{\varnothing}_t)=W_{r2}*RELU(W_{r1}*([\boldsymbol{\varnothing}_x,\boldsymbol{\varnothing}_t])) \tag{4}$$

式(4)表示在文本特征$\boldsymbol{\varnothing}_t$指导下，计算残差特征$f_{res}$。与式(2)和式(3)相似，*表示带批标准化的二维卷积操作，$RELU(\cdot)$表示Relu激活函数，W_{r1}和W_{r2}表示3×3的卷积核，[·]表示拼接操作。

TIRG模块与已有的特征融合方法有所区别。已有的方法通常是将文本特征和图片特征“平等”地融合到一起，产生一个新的联合特征；而TIRG方法侧重于用文本特征“修改”源图片特征，得到的组合特征仍存在于图片特征向量空间中。其中，门控操作将原始图片的特征作为输出组合特征的一个重要参考，使两者共存于同一个高维向量空间。另一方面，TIRG模块在门控特征上添加残差特征，实现了在图片向量空间中的“游走”(“修改”)操作。TIRG模块可以将文本的“修改”作用于图片特征提取器的全连接层(此时$W=H=1$)，也可以作用于最后的卷积层，我们将其视为一个超参数。

(3)损失函数设计

与普通的检索任务相似，TIRG模型的训练目标是使组合特征和目标图片特征(正例对)之间的距离尽可能小，与不相关的图片特征(负例对)之间的距离尽可能大，这里使用分类损失函数来训练模型。给定包含B个查询项的训练批次，使用$\boldsymbol{\psi}=f_{\text{combine}}(x_i^{\text{query}},t_i)$表示第$i$个查询项的特征(即源图片被文本修改后的组合特征)，用$\boldsymbol{\varnothing}_i^+=f_{\text{img}}(x_i^{\text{target}})$表示它对应的目标图片特征。对于第$i$个查询项特征$\boldsymbol{\psi}_i$，为它构造一个大小为$K$的集合$N_i$，其中包含一个正例(positive example)$\boldsymbol{\varnothing}_i^+$和$K-1$个负例$\{\boldsymbol{\varnothing}_1^-,\boldsymbol{\varnothing}_2^-,\cdots,\boldsymbol{\varnothing}_{K-1}^-\}$($K$小于或等于$B$)。重复这个过程$M$次(记作$N_i^m$)，来构造所有可能的集合，其中$M$的最大值是$C_B^K$(实验中，为了提升训练效率通常取一个更小的值)。我们用交叉熵损失函数来训练模型，如式(5)所示：

$$L=\frac{-1}{MB}\sum_{i=1}^{B}\sum_{m=1}^{M}\log\left\{\frac{\exp\{k(\boldsymbol{\psi}_i,\boldsymbol{\varnothing}_i^+)\}}{\sum\limits_{\boldsymbol{\varnothing}_j\in N_i^m}\exp\{k(\boldsymbol{\psi}_i,\boldsymbol{\varnothing}_j)\}}\right\} \tag{5}$$

其中，$k(\cdot)$表示计算相似度的函数，在实验中为计算内积。

3.2.2 选择题策略

我们设计了一种贪心搜索算法，利用TIRG组合式图片检索模型来完成美妆图片排序任务。已知TIRG模型可以根据一张初始的人脸图片和一系列步骤描述，检索出下一状态的人脸图片。我们先从5张乱序图片中挑选出1张，作为初始图片。挑选初始图片的方法有两种：1)穷举问题中的5张图片；2)基于第一个图片对比较模型，通过两两比较得到图片的相对顺序，挑选排在最前面的图片作为初始图片。随后进行贪心搜索算法的初始化：将挑选的初始图片作为源图片I_1'，默认给定问题中的第二个化妆步骤$\{S_k\}_{k=2}^2$为初始文本描述。接下来，通过贪心算法以最大化图片检索的匹配分数为目标，进行图片的迭代检索。其中，每一轮检索后都更新源图片和步骤描述，最终得到一个由初始图片和每轮检索图片构成的有序图片集$\{I_1',I_2',I_3',I_4',I_5'\}$，从而确定了5张图片的相对顺序。算法1展示了使用TIRG模型多轮迭代产生图片顺序的过程。

算法1 TIRG模型产生图片顺序的贪心搜索算法

输入：乱序的人脸图片$\{I_i\}_{i=1}^N$；有序步骤描述$\{S_j\}_{j=1}^M$；图片对比较模型$f_{\text{img}}(\cdot)$；TIRG模型$f_{\text{tirg}}(\cdot)$

输出：有序的人脸图片$\{I_i'\}_{i=1}^N$

1. 基于图片对比较模型挑选出初始图片：
   $I_i'=f_{\text{img}}(\{I_i\}_{i=1}^N)$；
2. 初始化源图片集$I_s=\{I_i'\}$，目标图片候选集$I_t=\{I_i\}_{i=1}^N-I_s$；
3. 图片索引 x=1，步骤索引 y=2；
4. while $|I_s|<N$ do
5. $I_{\text{target}}^x=\boldsymbol{\varnothing}$，$S_{\text{target}}^x=0$；
6. for (j=y; j<M−(N−x); j++) do
7. $I_j=\arg\max\limits_{I\in I_t} f_{\text{tirg}}(I_x',\{S_k\}_{k=y}^j,I)$；
8. $S_j=\max\limits_{I\in I_t} f_{\text{tirg}}(I_x',\{S_k\}_{k=y}^j,I)$；
9. if $S_j>S_{\text{target}}^x$ then
10. $I_{\text{target}}^x=I_j$，$S_{\text{target}}^x=S_j$；
11. else
12. y=j; break;
13. end if
14. end for
15. $I_{x+1}'=I_{\text{target}}^x$；$I_s=I_s+\{I_{x+1}'\}$；
16. $I_t=I_t-\{I_{x+1}'\}$；x=x+1；
17. end while

基于TIRG模型产生一组图片顺序，我们用最小编辑距离(Levenshtein距离)从候选答案中挑出最终的排序答案。最小编辑距离指两个序列之间，由一个转换到另一个需要的最小编辑操作(包括插入、删除、替换)次数。例如，答案序列[1,2,3,4,5]和[2,1,3,4,5]之间的最小编辑距离是1。

4 数据处理与分析

4.1 数据集介绍

YouMakeup VQA Challenge为美妆图片排序任务提供了YouMakeup数据集。它一共包含2 800个来自YouTube网站的高质量美妆教学视频，视频总时长超过420 h。其中每个视频都标注了所有化妆步骤的起始结束时间、作用的面部区域位置和对应的英文描述。平均每个视频包含了10.9个化妆步骤。

官方为图片排序任务提供了训练数据和评测数据。训练集一共包含1 680个视频，验证集一共包含280个视频，用于参赛者模型的训练。同时，由于图片排序任务以视觉问答的形式给出，因此官方在包含280个视频的验证集上产生了1 200个问题，在包含420个视频的测试集上产生了1 500个问题，用于排序任务的评测。在美妆图片排序任务上，YouMakeup数据集的统计信息如表1所列。

表1 YouMakeup数据集的统计信息

Table 1 Data statistic of YouMakeup

	视频数	平均步骤数	任务问题数
训练集	1 680	10.57	—
验证集	280	11.31	1 200
测试集	—	—	1 500

4.2 数据预处理

我们通过预处理官方提供的美妆视频数据，得到了训练

本文两个排序模型所需的图片-文本数据。我们将每个完整的视频按照标注的化妆步骤时间划分成小的视频片段，每个视频片段呈现一个化妆动作。然后从每个视频片段的结尾处抽取 10 帧，确保这些图片呈现了某个化妆步骤结束时人脸的一个形态，而不是步骤未完成时的中间态。细节上，如果一个视频片段的时长大于 10 s，从后往前每隔 5 帧抽 1 帧；如果时长小于 10 s，则连续抽帧。同时，保留每个化妆步骤的标注（作用的面部区域位置和对应的英文描述）作为对应图片的标注信息。基于 YouMakeup 视频数据集，一共处理得到 177 390 张训练集图片和 31 670 张验证集图片，其统计信息如表 2 所列。

表 2 构造的图片数据集统计信息

Table 2 Data Statistic of extracted images

	训练集	验证集
视频数	1 680	280
步骤数	17 739	3 167
抽取图片数	177 390	31 670

随后，用预先抽取的带标注美妆图片，构造两个模型的训练集和验证集。对于图片对比较模型，随机选取来自同一个视频的两个步骤，用它们对应的图片构成一个图片对（I_i，I_j）。一个包含 N 个步骤的视频，最多可以构造得到 C_N^2 个图片对。为了保证训练数据的平衡性，在构造数据的过程中设置一个随机数种子，使得属于类别 0 和类别 1 的数据各占 50%。对于组合式图片检索模型，随机地将来自同一个视频的化妆步骤分成 N 个部分，每个部分可以构造得到一个三元组（I_i，S_i，I_j），其中 I_i 表示源图片，S_i 是拼接第 i 部分所有步骤描述得到的文本，I_j 表示目标图片。关于两个模型的实验数据统计信息如表 3 所列。

表 3 两个模型的实验数据统计信息

Table 3 Data statistic of constructed training and validation data for two models

	训练集	验证集
步骤数	17 739	3 167
抽取图片数	177 390	31 670
图片对数	117 226	21 544
三元组数	333 780	12 114

4.3 数据分析

根据预处理得到的带标注图片数据，我们进一步分析了美妆领域数据的特点，得到了以下结论：

（1）不同化妆动作带来的视觉变化差异很大。在化妆过程中，不同的化妆步骤作用在同一张人脸上，即它们拥有相似的背景。但同时由于每个化妆步骤之间都有区别，它们表示不同的化妆动作，作用的人脸部位和带来人脸形态的改变也不近相同。有些化妆动作可以给人脸带来明显的变化，例如“在嘴唇上涂上红色口红”，而有些动作仅仅带来细微的，甚至难以察觉的变化，例如“在脸上用粉底刷涂上粉底液”只会使人脸的整体色调发生变化。

（2）眼部和面部皮肤是数据中需要最多化妆操作的部位，如图 5 所示。我们统计了所有化妆步骤作用的人脸部位，发现一共有 24 个部位。将其中相近的区域进行整合，例如将“人中”“前额”“脸颊”等整合到“面部皮肤”区域，最后得到五大类区域，即眼部、鼻子、眉毛、唇部、面部皮肤，并统计它们对应的操作次数，如图 5 所示。

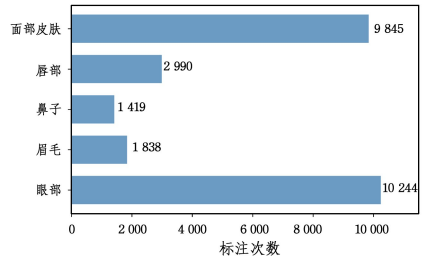


图 5 数据集中各人脸部位操作次数的统计

Fig. 5 Statistical data of operation times of each face area

（3）美妆图片中，人脸姿势变化和化妆品遮挡会产生干扰，如图 6 所示。在美妆图片排序任务中，我们应重点关注化妆品相关的视觉信息，主要体现在面部区域纹理、颜色、色调、形状等的变化，而这些变化和人脸姿势引起的形变、光照变化等耦合在一起，增加了排序的难度。同时，抽取的人脸图片包含化妆品部分，且存在严重遮挡的问题，也会引入噪声。

图 6 不同的人脸姿势和化妆品遮挡示例

Fig. 6 Examples of various face poses and cosmetic occlusions

（4）化妆步骤描述的句子结构有规律。一个典型的化妆步骤描述包含 3 个部分：化妆品、作用的人脸部位、化妆工具，例如“Apply foundationon the face with brush”。其中，作用的人脸部位信息能帮助定位到图片的重点区域，化妆品信息决定了人脸将会发生的变化，化妆工具可以作为一种辅助信息来对齐文本和人脸图片。

5 实验设置

5.1 评测指标

（1）分类准确率。图片对比较模型的本质是一个二分类器，因此使用准确率（Accuracy）作为评测指标之一。同时，由于该模型解决的是一个成对排序（Pair-wise rank）问题，默认标记排在前面的图片“排序等级”为 1，排在后面的图片“排序等级”为 0，因此使用 MRR（Mean Reciprocal Rank）和 nDCG（normalized Discounted Cumulative Gain）两个额外指标来更全面地衡量模型的表现。

（2）R@K。在组合式图片检索任务中，使用 R@K 来衡量检索的精度。它表示对于测试的所有查询项，满足正确的目标图片排在检索到的前 K 张图片的百分比。在美妆图片排序中，目标是能够在 5 张图片中实现精准的检索，我们取 $K=1,2,3$。此外，因为该任务的目标是对来自同一个视频不同步骤的图片进行排序，检索时与其他视频的图片无关，所以我们先计算每个视频中的 $R@K$ 值，再对所有视频取平均得到最终的 $R@K$ 值，以减少噪声。

（3）选择准确率。美妆图片排序任务以视觉问答的形式给出，官方为每个问题提供了 4 个候选答案，其中包含 1 个正确答案，因此采用选择准确率来衡量模型在此任务上的表现。

选择准确率等于选出正确答案的问题数占总问题数的比例。

5.2 实验细节

图片对比较模型的设置如下：使用经过 ImageNet[22] 预训练的 ResNet-18 网络作为孪生网络的图片特征提取器，取 ResNet-18 最后池化层的输出作为特征向量，特征维度是 512。在模型的训练中，本文使用了随机梯度下降法（Stochastic Gradient Descent，SGD）作为优化器，*batch size* 设定为 32，初始学习率设定为 0.0001。每训练 5 轮学习率缩小为原来的 1/10，共训练 8 轮。

组合式图片检索模型的设置如下：同样使用经过 ImageNet 预训练的 ResNet-18 网络作为图片编码器，输出 512 维的特征向量。文本采用单层的循环神经网络 LSTM 作为编码器，随机初始化网络参数。输入一个文本句子，先将其中的每个单词转换为小写，得到它们的 One-hot 向量表示，再将其转换为 512 维的词向量。如果一个单词在整个文本集合中出现的次数少于 5，则用特殊符号〈UNK〉代替。最后这些词向量经过 LSTM 编码得到 512 维的句子特征。我们将文本的修改作用于图片特征提取器的全连接层，交叉熵损失函数中取 $K=B, M=1$。在训练模型时，设定随机梯度下降法为优化器，设定 *batch size* 为 32，设定初始学习率为 0.01。每训练 12 轮学习率缩小为原来的 1/10，共训练 34 轮。

实验中所有代码的编写都基于深度学习框架 Pytorch[1)]。

6 实验结果和分析

在预处理得到的图片数据集上，我们分析了图片对比较模型（Pairwise）的分类准确率和组合式图片检索模型（TIRG）的 R@K 结果。进一步地，我们在两个模型上探究了图片间隔化妆步骤数对实验结果的影响。最后，我们比较并分析了两者在美妆图片排序任务上的选择准确率。

6.1 图片对比较模型的结果

图片对比较（Pairwise）模型在分类任务和排序任务上的实验结果如表 4 所列。作为一个二元分类器，它在测试集上取得了 82.65%的准确率。当将其看作一个 pairwise ranking 问题时，它的 *MRR* 和 *nDCG* 值分别达到了 0.9138 和 0.9367，同样表现不俗。相比随机分类 50%的分类准确率，Pairwise 模型在没有文本描述的指导下，也能在大多数情况下正确判断两张图片的相对顺序。这个实验结果也证实了一个直观的想法：化妆是一个不断往脸上“加”东西（化妆品）的过程，因此在大部分情况下，仅仅对比两张图片的信息，就能得到它们的相对顺序。在美妆图片排序任务上，Pairwise 模型在包含 1200 个问题的验证集上选择准确率为 65.70%，在包含 1500 个问题的测试集上选择准确率为 67.90%。通过对比发现，模型的选择准确率明显低于分类准确率。因为在图片排序任务中，Pairwise 模型需要比较 C_5^2 个图片对的相对顺序，才能确认 5 张图片的最终顺序。而在多轮图片对比中，模型容易被其中几个难分图片对误导，从而导致排序错误。由表 4 可以发现，Pairwise 模型虽然在两张图片的排序问题中有不俗的表现，但是无法很好地解决多张图片的排序问题，因此文本的额外指导是不可或缺的。

表 4 图片对比较模型（Pairwise）的实验结果

Table 4 Performance of image pairwise comparison model

	Accuracy/%	*MRR*	*nDCG*	Choice Accuracy/%	
				Valid	Test
Random	50.00	0.7500	0.8154	25.00	25.00
Pairwise	82.65	0.9138	0.9367	65.70	67.90
Pairwise+	83.65	0.9182	0.9399	66.20	70.00

我们进一步探究了两张图片间隔的化妆步骤数对分类准确率的影响。通过观察化妆过程，我们有一个直观的感受：一般情况下，当两张图片间隔的化妆步骤数越少时，人脸形态的变化也越小；当间隔的化妆步骤数越多时，人脸形态的变化也越大。例如，当只经过“画外眼线”一个化妆步骤时，眼部发生的变化几乎难以察觉；而当经过“画外眼线、画内眼线、夹睫毛、涂睫毛膏”等多个化妆步骤时，眼部的变化会趋于明显。我们按照间隔的化妆步骤数（记为 Step gap）重新划分验证集中的图片对，然后在每个子集上重新测试模型的分类准确率，实验结果如表 5 所列。结果显示，Pairwise 模型的分类准确率会随着 *Step gap* 的增加而提高，而其中最难区分的是间隔一个化妆步骤（*Step gap*=1）的图片对，从而验证了上述直观想法的合理性。

表 5 Pairwise 模型在不同 *Step gap* 上的分类准确率

Table 5 Classification accuracy on different *Step gap* for image pairwise comparison model

（单位：%）

Step gap	1	2	3	4
Pairwise	61.51	70.94	76.87	82.96
Pairwise+	61.90	71.59	78.95	84.33

考虑到 Step gap 会影响 Pairwise 模型的分类准确率，我们引入了课程学习[13]的策略来改进模型。先将训练集数据按 *Step gap*=[1,2,3,4,≥5]进一步划分为 5 个子集，将间隔步骤数小的图片对记为困难样例（hard samples），将间隔步骤数大的图片对记为简单样例（easy samples）。在模型训练中，我们对应划分了 5 个阶段，在每个阶段使用不同的数据由易到难地微调模型参数。将加入课程学习策略的图片对比较模型记做 Pairwise+，它的实验结果如表 4 末行所列，其在不同 *Step gap* 上的具体表现如表 5 末行所列。实验结果表明，课程学习策略能小幅度提升模型的分类准确率和选择准确率。

6.2 组合式图片检索模型的结果

组合式图片检索（TIRG）模型在分类任务和排序任务上的实验结果如表 6 所列。相比于随机检索的结果，TIRG 模型在验证集的 R@1，R@2，R@3 这 3 个指标上的表现都有大幅度的提升，分别达到了 30.15%，49.16%，63.20%。此外，我们还对比了纯图片检索（Image-to-Image）和 TIRG 模型的表现。其中，纯图片检索方法直接用源图片检索目标图片，得到的结果如表 6 第 3 行所列。对比两者的表现可以看到，TIRG 模型在验证集的 R@1，R@2，R@3 这 3 个指标上都有绝对优势。这说明直接通过计算两张图片的相似度来检索目标图片的效果较差，而 TIRG 在此基础上加入了修改文本的

1) https://github.com/AIM3-RUC/YoumakeupBaseline/tree/master/image_ordering

信息，确实能使得源图片特征和目标图片特征更加接近，从而提升检索的精度。

表 6 TIRG 模型的实验结果

Table 6 Performance of TIRG image retrieval model

	R@1	R@2	R@3	Choice Accuracy/%	
				Valid	Test
Random	9.55	19.11	28.66	25.00	25.00
TIRG	30.15	49.16	63.20	58.67	58.93
Image-to-Image	17.22	31.37	43.69	—	—

在 TIRG 模型中，我们同样探究了 Step gap 对模型检索表现的影响。类似地，Step gap 定义为源图片与目标图片之间间隔的步骤数，对应的修改文本则是图片间所有步骤描述的拼接。例如，对于三元组(I_i, S_k, I_j)，图片I_i是源图片，图片I_j是目标图片，修改文本$S_k = Concate[S_{i+1}, \cdots, S_j]$$(j \geqslant i+1)$，$Step\ gap = (j-i)$。我们将验证集中的三元组按照间隔的化妆步骤数(Step gap)重新划分，然后在每个子集上重新测试模型的 $R@K$，实验结果如表 7 所列。实验结果表明，与 Pairwise 模型的表现相反，TIRG 模型的检索精度随着 $Step\ gap$ 的增加而降低。我们认为主要有两个原因：1)随着 $Step\ gap$ 增大，修改的文本包含的句子增多，文本特征编码器更难捕捉句子中的上下文语义信息；2)$Step\ gap$ 越大，源图片与目标图片之间的差异越大，源图片特征需要在特征向量空间中“游走”更大的距离才能接近目标图片的特征，这对 TIRG 模型来说无疑是更难的。两个模型在 Step gap 上相反的表现说明了两者具有互补性，我们可以选择用 Pairwise 模型来预测 $Step\ gap$ 较大的图片对顺序，而用 $TIRG$ 模型来预测 $Step\ gap$ 较小的图片对顺序。

表 7 TIRG 模型在不同 $Step\ gap$ 上的 $R@K$

Table 7 $R@K$ on different $Step\ gap$ for TIRG model

$Step\ gap$	1	2	3	4
R@1	34.73	28.83	27.62	26.92
R@2	54.83	50.69	47.93	46.96
R@3	68.54	65.62	62.54	62.92

6.3 两个模型实验结果的对比

上文是对两个模型实验结果的单独分析，下面将分析和对比两个模型在任务选择题上的准确率。由表 4 和表 6 可以得到，改进后的 Pairwise 模型在测试集上的选择准确率可以达到 70%，而 TIRG 模型只能达到 58.93%，相差甚远。这个实验结果和预期不符，其中主要存在两个原因：1)两个模型在图片排序任务上选择答案的算法具有很大差异。Pairwise 模型只要通过两两对比即可以得到 5 张图片的最终顺序。而对于 TIRG 模型，每个问题中不定数目的化妆步骤和 5 张图片不具有对应关系，需要通过复杂的算法将化妆步骤分割成几部分以得到对应的修改文本，存在很大的噪声。2)TIRG 模型使用了粗粒度的图片和文本特征，不能有效地捕捉文本的指导信息。

本文进一步探究 TIRG 模型是否在一定程度上捕捉到了文本的语义指导信息。由于两个模型的评测指标不同从而无法直接比较，我们在这里转变了 TIRG 模型的评测方式，与 Pairwise 模型保持一致。具体地，对于三元组(I_i, S_k, I_j)，可以通过 TIRG 模型计算查询项特征 $f(I_i, S_k)$和目标图片特征 $f(I_j)$的相似度分数$score_{i\to j}$，同时反向计算得到相似度分数$score_{j\to i}$。如果$score_{i\to j} > score_{j\to i}$，说明图片 i 是源图片，出现的时间更早，对应类别为 1；反之图片 j 出现的时间更早，对应类别为 0。基于这种方式，我们可以用 TIRG 来计算分类准确率。

考虑到上述实验中两个模型的训练数据有差异，下文将严格保证两个模型训练数据的一致性，以进行更公平的对比。我们构造了 $Step\ gap = 1$ 的 78 375 个三元组数据(I_i, S_k, I_j)数据(I_i, S_k, I_j)，同时取出其中的(I_i, I_j)作为对应的图片对数据，分别用来重新训练两个模型。在包含 2 887 个三元组(图片对)的验证集上，我们得到了两个模型的分类准确率，如表 8 所列。从实验结果可以看到，在 $Step\ gap = 1$ 的数据上，TIRG 模型表现得比 Pairwise 略好，这说明 TIRG 模型确实能捕捉到文本的语义信息，但表现乏力。

表 8 Pairwise 模型和 TIRG 模型的分类准确率对比

Table 8 Comparison of classification accuracy between Pairwise and TIRG model

Model	$Accuracy$/%
Pairwise	61.24
TIRG	61.69

结束语 本文根据 YouMakeup VQA Challenge 探究了一个新颖多模态任务，即基于有序步骤描述，进行美妆图片顺序的推理。为了解决这个任务，一方面要有效融合文本和图片两种模态的数据，另一方面要从细粒度的角度来提升模型对文本和图片的语义理解能力。本文从单模态和多模态的角度提出了两个模型，实验结果表明，两个模型在美妆图片排序任务上有良好的表现，在测试集上的选择准确率分别为 70% 和 58.93%。而进一步的实验表明，虽然多模态模型在选择准确率上的表现劣于单模态模型，但其实质上捕捉到了文本的语义信息，并与单模态模型具有互补性。

未来的工作可以从 3 方面进行改进：1)在图片端，检测出不同的人脸区域和物体，追踪它们的连续变化；2)在文本端，探索更细粒度的特征来充分挖掘其指导作用，例如可以分解句子的结构；3)实现文本和图片之间更细粒度的交互，例如可以根据文本定位到具体的人脸区域，同时结合两个模型在 Step gap 上的互补性。

参 考 文 献

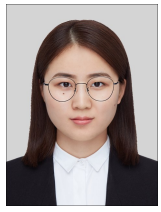
[1] CHEN S, WANG W, RUAN L, et al. YouMakeup VQA Challenge: Towards Fine-grained Action Understanding in Domain-Specific Videos[J]. arXiv: 2004.05573.

[2] TONG W S, TANG C K, BROWN M S, et al. Example-based cosmetic transfer [C]// 15th Pacific Conference on Computer Graphics and Applications (PG'07). IEEE, 2007: 211-218.

[3] GU Q, WANG G, CHIU M T, et al. Ladn: Local adversarial disentangling network for facial makeup and de-makeup [C]// IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019: 10480-10489.

[4] GUO D, SIM T. Digital face makeup by example [C]// IEEE Conference on Computer Vision and Pattern Recognition.

Miami, FL: IEEE, 2009: 73-79.
[5] CHEN H J, HUI K M, WANG S Y, et al. Beautyglow: On-demand makeup transfer framework with reversible generative network [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach. CA, USA: IEEE, 2019: 10034-10042.
[6] LI Y, HUANG H, YU J, et al. Cosmetic-Aware Makeup Cleanser[J]. arXiv: 2004.09147.
[7] WANG W, WANG Y, CHEN S, et al. YouMakeup: A Large-Scale Domain-Specific Multimodal Dataset for Fine-Grained Semantic Comprehension [C]// Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: ACL, 2019: 5136-5146.
[8] VO N, JIANG L, SUN C, et al. Composing text and image for image retrieval-an empirical odyssey [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 6432-6441.
[9] NAM H, HA J W, KIM J. Dual attention networks for multimodal reasoning and matching [C]// IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI: IEEE, 2017: 2156-2164.
[10] LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching [C]// European Conference on Computer Vision. Springer, Cham, 2018: 201-216.
[11] LI K, ZHANG Y, LI K, et al. Visual semantic reasoning for image-text matching [C]// International Conference on Computer Vision. IEEE, 2019: 4654-4662.
[12] CHEN H, DING G, LIN Z, et al. Cross-modal image-text retrieval with semantic consistency [C]// Proceedings of the 27th ACM International Conference on Multimedia. Nice, France, ACM, New York, NY, USA, 2019: 1749-1757.
[13] WANG T, XU X, YANG Y, et al. Matching images and text with multi-modal tensor fusion and re-ranking [C]// In Proceedings of the 27th ACM International Conference on Multimedia. ACM, 2019: 12-20.
[14] GUO X, WU H, CHENG Y, et al. Dialog-based interactive image retrieval [C]// Advances in Neural Information Processing Systems. MIT Press, 2018: 678-688.
[15] HOSSEINZADEH M, WANG Y. Composed Query Image Retrieval Using Locally Bounded Features [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 3596-3605.
[16] PARK D H, DARRELL T, ROHRBACH A. Robust change captioning[C]// IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019: 4623-4632.
[17] TAN H, DERNONCOURT F, LIN Z, et al. Expressing visual relationships via language[J]. arXiv: 1906.07689.
[18] BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning [C]// Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada: ACM, 2009: 41-48.
[19] CHOPRA S, HADSELL R, LECUN Y. Learning a similarity metric discriminatively, with application to face verification [C]// Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005: 539-546.
[20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 770-778.
[21] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
[22] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [C]// In Advances in Neural Information Processing Systems. MIT Press, 2012: 1097-1105.



YAO Lin-li, born in 1998, postgraduate. Her main research interests include image-text matching and visual semantic understanding.

JIN Qin, born in 1972, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include multimedia computing and human computer interaction.