https://crad.ict.ac.cn
62(3): 662−671, 2025

基于对比学习的跨模态实体链接模型

王苑铮　孙文祥　范意兴　廖华明　郭嘉丰
（中国科学院网络数据科学与技术重点实验室（中国科学院计算技术研究所）　北京　100190）
（中国科学院大学　北京　100049）
（wangyuanzheng19z@ict.ac.cn; ）

A Cross-Modal Entity Linking Model Based on Contrastive Learning

Wang Yuanzheng, Sun Wenxiang, Fan Yixing, Liao Huaming, and Guo Jiafeng
（*CAS Key Laboratory of Network Data Science & Technology* (*Institute of Computing Technology*, *Chinese Academy of Sciences*), *Beijing* 100190）
（*University of Chinese Academy of Science*, *Beijing* 100049）

Abstract　Image-text cross-modal entity linking is an extension of traditional named entity linking. The inputs are images containing entities, which are linked to textual entities in the knowledge base. Existing models usually adopt a dual-encoder architecture which encodes entities of visual and textual modality into separate vectors, then calculates their similarities using dot product, and links the image entities to the most similar text entities. The training process usually adopts the cross-modal contrastive learning task. For a given modality of entity vectors, this task pulls closer the vector of another modality that corresponds to itself, and pushes away the vector of another modality corresponding to other entities. However, this approach overlooks the differences in representation difficulty within the two modalities: visually similar entities are often more difficult to distinguish than textual similar entities, resulting in the incorrect linking of the former ones. To solve this problem, we propose two new contrastive learning tasks, which can enhance the discriminative power of the vectors. The first is self-contrastive learning, which aims to improve the distinction between visual vectors. The second is hard-negative contrastive learning, which helps a textual vectors to distinguish similar visual vectors. We conduct experiments on the open-source dataset WikiPerson. With a knowledge base of 120 000 entities, our model achieves an accuracy improvement of 4.5% compared with the previous state-of-the-art model.

Key words　entity linking model; multi-modal; cross-modal; contrastive learning; visual information

摘　要　图文跨模态实体链接是对传统实体链接任务的扩展，其输入为包含实体的图像，目标是将其链接到文本模态的知识库实体上. 现有模型通常采用双编码器架构，将图像、文本模态的实体分别编码为向量，利用点乘计算两者的相似度，从而链接到与图像实体相似度最高的文本实体. 其训练过程通常采用基于InfoNCE损失的对比学习任务，即提高一个实体某模态与自身另一模态的向量相似度，降低与其他实体另一模态的向量相似度. 然而此模型忽视了图文2个模态内部表示难度的差异：图像模态中的相似实体，通

收稿日期：2023-09-12；**修回日期**：2024-07-02
基金项目：国家自然科学基金项目（62372431）；国家重点研发计划项目（2021QY1701, 2023YFA1011602）；中国科学院青年创新促进会会员项目（2021100）；中国科学院计算技术研究所创新项目（E261090）；国防科技创新项目
This work was supported by the National Natural Science Foundation of China (62372431), the National Key Research and Development Program of China (2021QY1701, 2023YFA1011602), the Youth Innovation Promotion Association Member Project of Chinese Academy of Sciences (2021100), the Innovation Project of Institute of Computing Technology of Chinese Academy of Sciences (E261090), and the National Defense Science and Technology Innovation Zone Foundation of China.
通信作者：范意兴（fanyixing@ict.ac.cn）

常比文本模态中的相似实体更难以区分，导致外观相似的图像实体很容易链接错误. 因此，提出 2 个新的对比学习任务来提升向量的判别能力. 一个是自对比学习，用于提升图像向量之间的区分度；另一个是难负例对比学习，让文本向量更容易区分几个相似的图像向量. 在开源数据集 WikiPerson 上进行实验，在 12 万规模的实体库上，相比于采用 InfoNCE 损失的最佳基线模型，模型正确率提升了 4.5 个百分点.

关键词 实体链接模型；多模态；跨模态；对比学习；视觉信息

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202330731 **CSTR:** 32373.14.issn1000-1239.202330731

实体链接(entity linking, EL)是一项经典的自然语言处理任务，其目标是将自然语言文本中有歧义的命名实体映射到知识库中相应的实体. 近年来，实体链接任务开始被扩展到文本以外的其他模态数据上. 其中一种新任务是图文跨模态实体链接任务[1]. 该任务的示例如图 1 所示，其输入是一张包含命名实体(例如图中方框标注的人物)的图像，而知识库中的实体是文本模态，包含名称、描述等信息，任务目标是将输入的图像实体正确链接到相应的文本实体上. 该任务对于多模态事实验证[2]、新闻图像字幕[3]等下游任务具有重要意义.

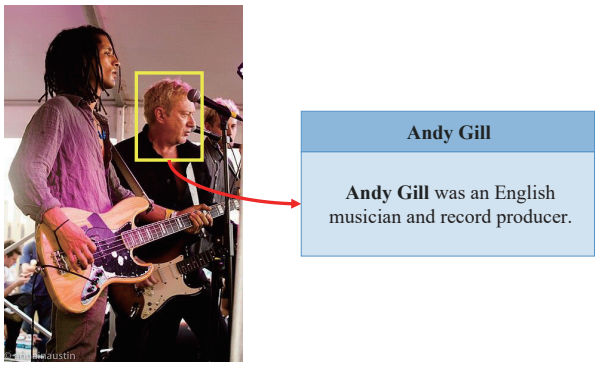


Fig. 1 An example of cross-modal entity linking task

图 1 跨模态实体链接任务实例

对于图文跨模态实体链接任务，当前模型通常采用双编码器架构. 此类模型利用 2 个独立的编码器，分别将图像实体、文本实体编码为向量表示，之后利用点乘计算两者的相似度，召回最相似的实体. 双编码器架构具有强大的向量化语义表示能力，可以将 2 种不同模态的实体编码为同一个语义空间下的向量，从而实现跨模态的语义对齐. 为了训练双编码器模型，主流方法采用基于 InfoNCE 损失的跨模态对比学习任务[1]. 其核心思想是对于一个实体某一模态的向量，提高它与自己另一模态向量的相似度，并降低它与其他实体另一模态向量的相似度. 该任务形式简洁，且实验验证其在跨模态实体链接任务上有效.

然而，InfoNCE 损失对图文 2 种模态一视同仁，忽视了图像、文本 2 种模态本身的差异：图像模态的相似实体，通常比文本模态的相似实体更难以区分. 对于文本实体，其通常包括姓名、性别、职业、国籍等清晰的描述性信息，即使相似的实体通常也会有一些信息不完全相同，因此能较容易地区分两者. 但图像实体有价值的信息较少，例如人脸、着装等，而这些有用信息又容易淹没在无用的背景信息里. 由于 2 个模态的差异，导致文本模态的实体向量区分度较大，而图像模态的实体向量区分度较小，反映为外观相似的图像实体容易彼此混淆，链接到错误的文本实体上.

针对以上问题，本文提出了 2 个新的对比学习任务，用于改进原本的 InfoNCE 损失. 一个是在图到文的匹配中引入自对比学习，通过降低不同图像实体的向量相似度，使其更容易彼此区分；另一个是在文到图的匹配中引入难负例(hard-negatives)对比学习，通过挖掘与输入相似的难负例图像，来提高一个文本实体对多个外观相似的图像实体的区分度. 这 2 个任务提升了向量的表示能力. 本文在针对人物实体的开源数据集 WikiPerson 上进行实验，其中在 12 万规模的实体库上，相比采用 InfoNCE 损失的最佳基线模型，本文的模型正确率提升了 4.5 个百分点，验证了本文模型的有效性.

1 相关工作

1.1 实体链接

实体链接是一项已被广泛研究的自然语言处理任务. 传统的实体链接任务，输入与实体库都是文本模态. 其模型的发展可以分为 3 个阶段：

1)基于规则的模型. 早期的实体链接技术主要采用基于规则的模型. 这种模型基于人工制定的规则，从文本中抽取实体，并与知识库中的实体进行匹配，代表性的工作有 OKKAM[4] 框架等. 由于制定规

则需要耗费大量的人力和时间, 因此这类模型的可扩展性较差. 此外, 这类模型难以处理实体名称的多样性等问题.

2)基于传统机器学习的模型. 机器学习模型相对于基于规则的模型, 更加灵活且性能更高, 因此逐渐取代了后者. 机器学习模型从实体提及以及上下文中提取特征, 如实体的类型、上下文词共现、实体属性等, 并训练机器学习模型做出链接决策. Milne等人[5]提出了WikipediaMiner模型, 首先对文本进行分词、词性标注、句法分析, 然后抽取实体上下文特征, 最后通过支持向量机等机器学习算法对实体进行分类和链接. Moro等人[6]基于图嵌入提出了Babelfy模型, 使用图嵌入技术将实体链接转换为图嵌入问题, 并通过模拟退火等算法对实体进行链接. 该模型可以自适应地学习实体链接任务中的特征和规则, 并具有很强的泛化能力和适应性. Ferragina等人[7]基于标签传播提出了Tagme模型, 将实体链接转化为图上的标签传播问题, 并通过迭代算法对实体进行链接. 该模型可以通过学习和优化实体上下文和链接概率来提高实体链接的准确性和鲁棒性.

3)基于深度学习的模型. 由于神经网络擅长建模抽象的语义, 在针对文本、图像等非结构化数据的任务上, 逐渐代替了传统机器学习模型. 在实体链接任务上, 此类模型利用神经网络建模上下文和实体之间的复杂语义关系. 尤其近年来, 借助BERT[8]等预训练语言模型强大的语义理解能力, 目前文本模态的深度学习实体链接模型已具有很高的正确性. Yamada等人[9]通过大规模语料预训练实体的向量表示, 用BERT编码局部和全局的实体信息来提高命名实体链接的准确率. Cao等人[10]则将命名实体链接转化成序列到序列的自回归任务, 输入包含命名实体的上下文, 并且直接解码出目标实体的标题. 如文献[10]所述, 这些基于预训练语言模型的实体链接模型, 已经接近了传统实体链接任务的性能上限. 因此近年来逐渐衍生出一些更具挑战性的实体链接任务, 例如零样本(zero-shot)链接[11−12]、跨模态链接[1]等.

1.2 多模态实体链接

传统实体链接任务中, 输入和知识库实体都是文本模态的. 而多模态实体链接任务是对这一设定的扩展, 它允许输入及知识库中包含文本之外的其他模态信息, 而目前的研究集中于添加图像模态.

根据输入上下文、输入中的待链接实体、知识库中的实体三者的模态组合方式, 当前的多模态实体链接可细分为4种任务设定, 如表1所示. 第1种设定中[13−17], 输入上下文为图文对(例如带图像的推特文本), 旨在将文本中的命名实体链接至知识库中文本模态的实体(通常包括名称、描述等信息). 其中输入中的图像起辅助作用. 第2种设定[17−22], 输入侧与第1种设定相同, 即利用图像和文本上下文, 链接文本中的命名实体, 不同之处在于知识库中的每个实体同时包含文本和图像信息(例如实体的图片). 第3种设定中[23], 虽然输入上下文仍为图文对, 但待链接的实体既可出现在文本中, 也可出现在图像中, 需将其分别链接到知识库中的文本或图像实体上. 上述3种设定的共同特点包括: 1)输入均为图文对, 允许利用2种模态的上下文信息; 2)待链接实体与知识库实体至少具有1个相同的模态. 换言之, 这3种设定均在单模态实体链接的基础上, 引入额外模态信息作为辅助.

Table 1 Task Settings and Representative Work of Multimodal Entity Linking

表1 多模态实体链接的任务设定与代表工作

输入		知识库实体	代表工作
上下文	实体		
文本+图像	文本	文本	文献[13−17]
文本+图像	文本	文本+图像	文献[17−22]
文本+图像	文本	文本	文献[23]
	图像	图像	
图像	图像	文本	文献[1]、本文

与前3种设定不同, 第4种设定下[1], 输入仅为包含实体的单张图像, 而知识库实体仅包含文本信息, 此设定要求在图像与文本这2种不同模态之间完成链接, 缺乏共同模态作为中介, 故称之为“跨模态实体链接”. 本研究即针对此设定进行.

1.3 多模态技术

互联网数据中, 有大量非结构化的文本、图像、视频、音频等不同模态的数据. 打通不同模态数据之间的壁垒, 可以挖掘出更多价值. 例如文献[24−25]发现图像和文本的信息可以相互补充与对齐.

借助于神经网络强大的语义表示能力, 原本不同模态的数据, 可以被编码为相同的表示(representation), 从而完成语义对齐. 其中一种常见方式, 是将所有模态的数据统一编码为向量表示. 早期的多模态技术通常是用多个异质网络单独编码每个模态的数据[26−29], 例如卷积神经网络(convolutional neural network, CNN)[30]编码图像数据, 长短期记忆网络(long short-term memory, LSTM)[31]编码文本数据等. 而较新的

技术则多采用基于自注意力的 Transformer[32] 模型. Transformer 最早被应用于处理文本数据，此后在视觉数据上，又产生了 ViT(vision transformer)[33] 等变种. 借助该结构，各模态数据的编码方式进一步统一.

在近年的多模态学习工作中，CLIP[34] (contrastive language-image pre-training)是一项集大成性的工作. 该模型用 RoBERTa[35] 作为文本编码器，用 ViT[33] 作为图像编码器，并且爬取了 4 亿条图文对齐的互联网数据，采用对比学习进行超大规模的预训练. CLIP 在很多任务上都达到了当前最高水平，成为了很多多模态模型的基础组件. 本文的模型也基于 CLIP 架构.

2 模　　型

本节首先形式化介绍图文实体链接任务，以及基于 InfoNCE 损失的对比学习建模方法. 之后介绍 2 个额外的任务：自对比学习、难负例对比学习. 通过引入这个 2 个任务，改进原始的 InfoNCE 损失，本文的模型可以学习到表示能力更强的向量.

2.1 基于 InfoNCE 损失的对比学习

在图文实体链接任务中，输入为包含 1 个或多个实体的图像 Img，称为图像实体；而知识库 $KB=\{Txt_1,Txt_2,\cdots,Txt_N\}$包含 N 个文本形态的实体 Txt_1, $Txt_2,\cdots,Txt_N$，称为文本实体，通常包含实体的名称、描述等文本信息. 实体链接任务，即检索 Img 对应的 $Txt\in KB$. 本文仅考虑目标实体在 KB 中的情况.

为了完成该任务，通常会采用双编码器架构，其结构如图 2 所示. 该结构包含 2 个独立的编码器，分别用来将图像实体、文本实体编码为向量，之后使用点乘计算图像、文本向量之间的相似度，并选择与输入图像相似度最大的文本实体作为链接目标. 此前的模型通常采用 InfoNCE 损失训练双编码器. 假设一个批次包含 N 个输入 $(Img_1, Txt_1), (Img_2, Txt_2),\cdots, (Img_N, Txt_N)$，其中 (Img_x, Txt_x) 代表 1 对匹配的图像、文本实体. 该模型首先用 2 个编码器将图像、文本实体编码为向量：

$$\boldsymbol{I}_x = ImageEncoder(Img_x), \tag{1}$$

$$\boldsymbol{T}_x = TextEncoder(Txt_x), \tag{2}$$

其中 $ImageEncoder$ 和 $TextEncoder$ 分别为图像、文本编码器，$\boldsymbol{I}_x$ 和 $\boldsymbol{T}_x$ 分别为编码后的图像、文本实体向量. 之后 InfoNCE 损失由 2 个子部分组成.

首先是图像到文本模态的对比学习损失 L^{I}：

$$L^{\mathrm{I}} = -\frac{1}{N}\sum_{x=1}^{N}\ln\frac{\exp(\boldsymbol{I}_x\cdot\boldsymbol{T}_x/\tau)}{\sum_{i=1}^{N}\exp(\boldsymbol{I}_x\cdot\boldsymbol{T}_i/\tau)}, \tag{3}$$

L^{I}利用点乘计算 $\boldsymbol{I}_x$ 与 $\boldsymbol{T}_i$ 之间的相似度，超参数 τ 为温度系数. 该任务用于学习图像到文本的匹配，其原理如图 3(a) 所示，L^{I} 旨在提高实体 x 的图像向量 $\boldsymbol{I}_x$ 与自身的文本向量 $\boldsymbol{T}_x$ 的相似度，并减小与其他实体的文本向量 $\boldsymbol{T}_i$ 的相似度.

另一个是文本到图像模态的对比学习损失 L^{T}：

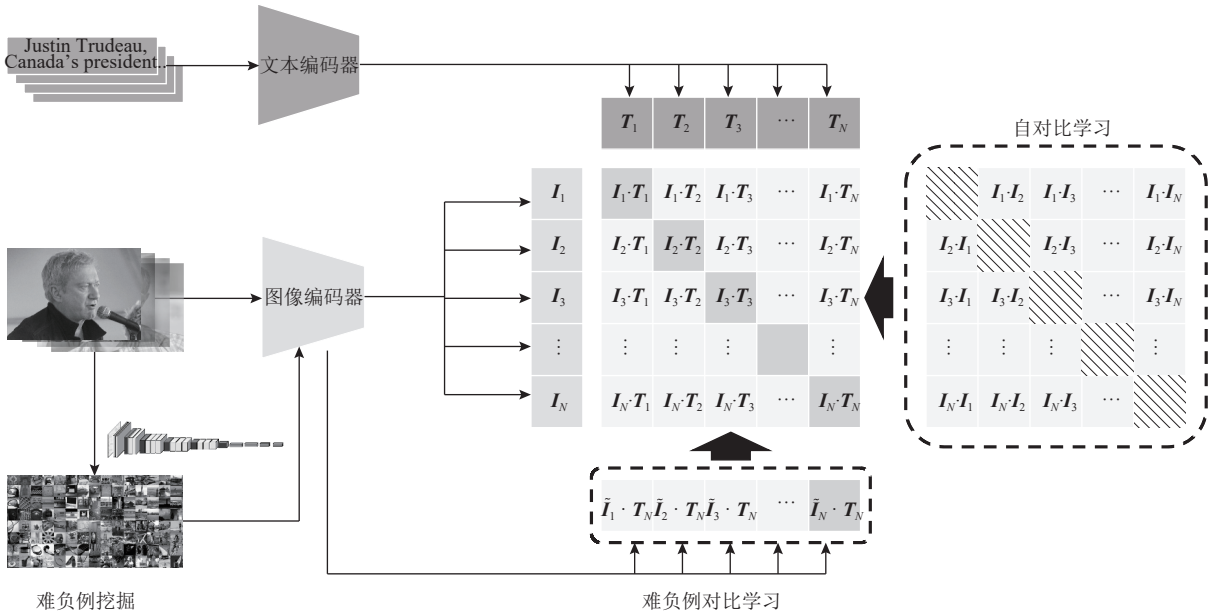
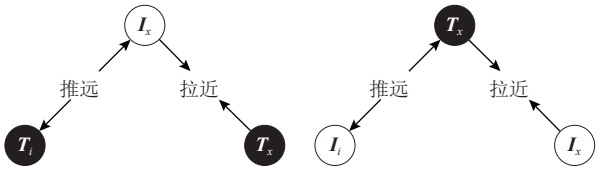


Fig. 2　Illustration of our model

图 2　本文模型示意图



(a) L^{I} (b) L^{T}

Fig. 3 Illustration of the InfoNCE loss

图 3 InfoNCE 损失示意图

$$L^{\mathrm{T}}=-\frac{1}{N}\sum_{x=1}^{N}\ln\frac{\exp(\boldsymbol{T}_x\cdot\boldsymbol{I}_x/\tau)}{\sum_{i=1}^{N}\exp(\boldsymbol{T}_x\cdot\boldsymbol{I}_i/\tau)}. \tag{4}$$

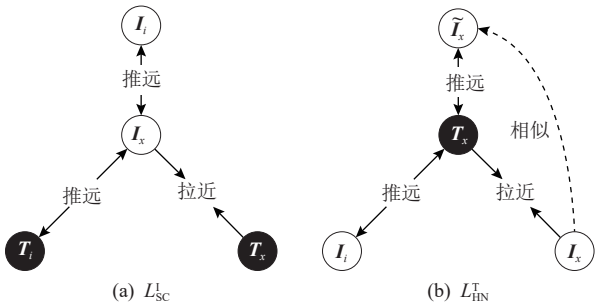
L^{T} 的形式与 L^{I} 对称, 其原理如图 3(b) 所示, 旨在提高实体 x 的文本向量 $\boldsymbol{T}_x$ 与自身的图像向量 $\boldsymbol{I}_x$ 的相似度, 并减小与其他实体的图像向量 $\boldsymbol{I}_i$ 的相似度.

最终的 InfoNCE 损失 L^{InfoNCE} 是 2 个函数的平均:

$$L^{\mathrm{InfoNCE}}=\frac{1}{2}\left(L^{\mathrm{I}}+L^{\mathrm{T}}\right). \tag{5}$$

由于 L^{I} 和 L^{T} 形式上的对称性, InfoNCE 损失对于图像、文本模态一视同仁. 该损失虽然简单易用, 却忽视了图像、文本模态表示难度的差异, 即图像模态的相似实体比文本模态更容易混淆, 因此限制了向量的表达能力.

为了解决这一问题, 本文在原本的对比学习任务中添加了自对比学习、难负例对比学习 2 个任务, 如图 4 所示. 如此得到了 2 个新的损失函数 $L^{\mathrm{I}}_{\mathrm{SC}}$ 和 $L^{\mathrm{T}}_{\mathrm{HN}}$, 分别用于替换 L^{I} 和 L^{T}, 从而提升向量的表达能力. 以下将详细介绍这 2 个任务.



(a) $L^{\mathrm{I}}_{\mathrm{SC}}$ (b) $L^{\mathrm{T}}_{\mathrm{HN}}$

Fig. 4 Illustration of our improved loss functions

图 4 本文改进后的损失函数示意图

2.2 自对比学习

自对比学习是一种无监督学习任务, 旨在通过分析样本之间的相似性与差异性, 有效学习样本的向量表示[36−38]. 其核心思想在于通过增大同一样本不同视图下的相似度, 以及减小不同样本的相似度, 来学习样本的表示. 以典型模型 SimCLR[37] 为例, 对于一个批次的图像样本, 该模型采取随机裁剪和颜色抖动等扰动操作, 产生不同视图的正例样本, 而将同一批次中的其他图像作为负例样本. 通过极大化正例图像间的相似度, 以及极小化负例图像间的相似度, 可以学习到表达能力强大的图像向量.

受此启发, 本文在图像到文本模态的匹配中, 添加了自对比学习(self-contrastive learning)任务. 对于图像实体 $\boldsymbol{I}_x$, 其正例为对应的文本实体 $\boldsymbol{T}_x$, 而负例为同一批次的其他图像实体 $\boldsymbol{I}_i$ 和文本实体 $\boldsymbol{T}_i$. 相应的损失函数 $L^{\mathrm{I}}_{\mathrm{SC}}$ 为

$$L^{\mathrm{I}}_{\mathrm{SC}}=-\frac{1}{N}\sum_{x=1}^{N}\ln\frac{\exp(\boldsymbol{I}_x\cdot\boldsymbol{T}_x/\tau)}{\sum_{i=1}^{N}\exp(\boldsymbol{I}_x\cdot\boldsymbol{T}_i/\tau)+\alpha_x}, \tag{6}$$

$$\alpha_x=\sum_{i\neq x}\exp(\boldsymbol{I}_x\cdot\boldsymbol{I}_i/\tau). \tag{7}$$

$L^{\mathrm{I}}_{\mathrm{SC}}$ 用于代替原本的 L^{I}. 其原理如图 4(a)所示, 与 L^{I} 相比, $L^{\mathrm{I}}_{\mathrm{SC}}$ 的分母额外增加了 α_x 项, 代表 $\boldsymbol{I}_x$ 不止要减小与其他文本向量 $\boldsymbol{T}_i$ 的相似度, 还要减小与其他图像向量 $\boldsymbol{I}_i$ 的相似度. 这一改进有利于让图像向量尽量彼此分隔开, 从而提升其表示能力.

2.3 难负例对比学习

在原本的对比学习任务中, 负例来自于一个批次中随机采样的其他样本. 然而这样的随机负例可能过于简单, 模型轻易就能区分开. 而实际情况下, 经常会存在相似的、难以判断的样本. 如果模型在训练时没有见过这些相似的样本, 则很难进行区分, 难以做出正确判断.

本文在文本到图像模态的匹配中, 引入了难负例对比学习. 这里的难负例, 指的是与输入图像相似的其他图像实体. 为了挖掘难负例, 本文首先在大规模人脸数据集 VGGFace2[39] 上训练了一个 ResNet50[40] 模型, 用于计算不同图像的人脸相似度. 该过程的损失函数为三元组损失, 其形式为

$$\sum_{i=1}^{N}\left[\|f(x_i^a)-f(x_i^p)\|^2-\|f(x_i^a)-f(x_i^n)\|^2+\gamma\right]_+, \tag{8}$$

其中 x_i^a 和 x_i^p 是同一人物的不同照片, 而 x_i^n 是与 x_i^a 不同人物的照片, $f(x)$ 是将人脸图像编码为向量的编码器, $[s]_+=\max(s,0)$, γ 为控制正负例距离的超参数. 该损失旨在让不同人脸之间的距离, 比同一人脸内部的 L2 距离至少大 γ.

ResNet50 训练完成后, 用其将实体链接数据集中的所有图像编码为向量, 召回与每个输入图像向量 $\boldsymbol{I}_x$ 欧氏距离最近但属于其他实体的图像 $\tilde{\boldsymbol{I}}_x$ 作为难负例.

在得到难负例图像$\tilde{\boldsymbol{I}}_x$后，将其加入到文本对图像的对比学习损失中，得到新的损失函数$L_{\mathrm{HN}}^{\mathrm{T}}$：

$$L_{\mathrm{HN}}^{\mathrm{T}}=-\frac{1}{N}\sum_{x=1}^{N}\ln\frac{\exp(\boldsymbol{T}_x\cdot\boldsymbol{I}_x/\tau)}{\sum_{i=1}^{N}\exp(\boldsymbol{T}_x\cdot\boldsymbol{I}_i/\tau)+\beta_x},\tag{9}$$

$$\beta_x=\exp\left(\boldsymbol{T}_x\cdot\tilde{\boldsymbol{I}}_x/\tau\right).\tag{10}$$

$L_{\mathrm{HN}}^{\mathrm{T}}$用于代替原本的$L^{\mathrm{T}}$. 其原理如图 4(b)所示，与$L^{\mathrm{T}}$相比，$L_{\mathrm{HN}}^{\mathrm{T}}$的分母额外增加了$\beta_x$项，代表$\boldsymbol{T}_x$不止要减小与批次内其他图像向量$\boldsymbol{I}_i$的相似度，还要减小与$\boldsymbol{I}_x$最接近的难负例$\tilde{\boldsymbol{I}}_x$的相似度. 这一改进可以让一个文本实体学会区分多个相似的图像实体，从而提升向量的判别能力.

最终的损失函数L为$L_{\mathrm{SC}}^{\mathrm{I}}$和$L_{\mathrm{HN}}^{\mathrm{T}}$的平均：

$$L=\frac{1}{2}\left(L_{\mathrm{SC}}^{\mathrm{I}}+L_{\mathrm{HN}}^{\mathrm{T}}\right).\tag{11}$$

此外本文也尝试了在文本到图像的匹配中添加自对比学习(降低不同实体间文本向量的相似度)，以及在图像到文本的匹配中添加难负例(挖掘与目标文本实体相似的文本，降低其与输入图像实体的相似度)，但是对性能没有明显提升. 其原因是文本模态的实体区分度比图像模态更大，即使不做额外处理，文本实体向量也不容易彼此混淆. 因此为了简化模型，本文没有添加这 2 项改进.

3 实验设置

3.1 数据集

本文在开源图文实体链接数据集 WikiPerson[1] 上进行了实验. 该数据集针对于人物实体，每张图像都包含 1 个或者多个待链接的人物，且人工标注了这些人物对应的维基百科实体. 该数据集包括共计 4.8 万张图像，全部来源于主流新闻网站. 其目标覆盖了超过 1.3 万个人物实体，其中包含了不同的职业、话题、肤色、年龄等. 由于绝大多数样本都只包含了 1 个目标人物，因此本文只考虑这类样本，并将包含多个目标人物的情况留待未来研究.

本文沿用了 Sun 等人[1] 的实验设置，为了观察模型在不同规模实体库上的表现，构建了 2 个不同大小的实体库，分别包含英文维基最高频的前 1.3 万实体和前 12 万实体. 在 1.3 万规模下，本文去掉了目标实体不在其中的样本，而 12 万规模可以覆盖全部目标实体. 在每种设置下，按照 8∶1∶1 的比例分为训练集、验证集、测试集.

3.2 基线模型

如 1.2 节和表 1 所述，尽管当前已有一些针对多模态实体链接的研究，但其中大部分并不适用于跨模态实体链接的场景. 这些研究与本文的主要区别在于：1)这些研究的输入为图文对，而本文的输入仅有图像；2)这些研究中，待链接实体与知识库实体有相同的模态，而本文中两者分属不同模态. 因此，大多数现有模型不适合与本文模型进行比较，而仅有 Sun 等人[1] 的设定与本文相同. 因此，本文对比的基线模型为：1)CLIP-Zero，即未经微调的 CLIP 模型[34]，直接在该数据集上进行零样本(zero-shot)测试；2)CLIP-InfoNCE，是 Sun 等人[1] 基于 InfoNCE 损失微调的 CLIP 模型，同时也是该任务上性能最好的基线模型.

3.3 模型设置

本文与 CLIP-InfoNCE 都采用 CLIP 作为双编码器架构，并训练了 2 个模型：1)CLIP-SC-HN，采用了$L_{\mathrm{SC}}^{\mathrm{I}}$和$L_{\mathrm{HN}}^{\mathrm{T}}$两个损失；2)CLIP-SC，未使用难负例对比学习，采用$L_{\mathrm{SC}}^{\mathrm{I}}$和$L^{\mathrm{T}}$两个损失.

本文采用 AdamW[41] 优化器训练模型，设置 ResNet 的学习率为 2E−4，CLIP 的学习率为 2E−6，这 2 个模型的批大小都为 64，训练 20 轮次. 所有损失函数中τ=0.07. 实验设备为 2 块 NVIDIA Tesla V100 GPU，采用了 5 组随机种子进行实验，并报告了 5 次实验的性能均值.

为了加快检索速度，在模型训练完成后，预先将所有文本实体编码为向量并存储起来，每次检索时只需编码图像实体的向量即可. 本文用稠密向量检索工具 FAISS[42] 检索文本向量，每个查询需要约 500 ms.

3.4 评价指标

本文采用了 2 个应用广泛的信息检索评价指标：Recall@K 和 MRR@K(mean reciprocal rank). 具体而言，本文报告了 Recall@1(也称为正确率 Accuracy)，Recall@10，MRR@10.

4 实验结果

4.1 主实验

表 2 展示了 4 个模型在 2 种规模的实体库上的性能. 从表 2 中可以观察到 5 个结论：

1)CLIP-Zero 虽然没有在图文实体链接数据集上微调，但仍有接近或超过一半实体能链接正确. 该现象表明，CLIP-Zero 在预训练阶段使用的互联网图文对齐数据集中，可能已经包括了一些常见人物实体，

Table 2 Performance of Models

表 2 模型性能 %

评测指标	模型	实体库规模	
		1.3 万	12 万
Accuracy	CLIP-Zero[34]	54.33	45.53
	CLIP-InfoNCE[1]	64.03	53.87
	CLIP-SC（本文）	64.81	56.96
	CLIP-SC-HN（本文）	**66.21**	**58.43**
Recall@10	CLIP-Zero	78.37	67.74
	CLIP-InfoNCE	86.87	76.75
	CLIP-SC（本文）	87.30	77.97
	CLIP-SC-HN（本文）	**88.04**	**79.14**
MRR@10	CLIP-Zero	62.28	52.79
	CLIP-InfoNCE	71.78	61.36
	CLIP-SC（本文）	72.40	63.92
	CLIP-SC-HN（本文）	**73.62**	**65.25**

注：黑体数值表示最优结果.

且这些实体信息已被模型成功地编码并存储.

2)CLIP-InfoNCE 的各项指标都明显优于 CLIP-Zero, 说明该模型在预训练中编码的实体知识是有限的, 不足以完成该任务, 利用 InfoNCE 损失进行微调可以有效提高图文实体的匹配性能.

3)CLIP-SC 的正确率相比于 CLIP-InfoNCE, 在小规模实体库上提升不明显, 而在大规模实体库上较为明显. 具体而言, 在 1.3 万实体库上提升了约 0.8 个百分点, 而在 12 万实体库上提升了约 3.1 个百分点. 这是因为自对比学习是一个仅针对图像模态的无监督任务, 对实体链接的影响是间接的. 当实体库规模较小时, 图像实体分布已经较为稀疏, 区分度较大, 即使继续增大它们之间的间隔也不会有明显增益. 而当实体库规模较大时, 图像实体分布较为稠密, 这时增大它们的间隔能明显改善向量的区分度, 从而提升链接性能.

4)CLIP-SC-HN 的正确率相比于 CLIP-SC, 在大规模和小规模实体库上的提升是相近的. 具体而言, 在 1.3 万实体库上提升了约 1.4 个百分点, 而在 12 万实体库上提升了约 1.5 个百分点. 相比于无监督的自对比学习, 难负例对比学习是有监督任务. 该任务可以直接改善文本实体对几个相似但不同的图像实体的判别能力, 在不同大小的实体库上都是有益的.

5)对于所有模型, 在大规模实体库上的性能明显差于小规模实体库. 因为实体库越大, 相似的实体就越多, 就越难以区分, 为学习实体表示带来很大挑战. 而相对于基线模型, 本文模型在实体库规模扩大后性能损失更小. 以 Accuracy 为例, 各个模型的性能下降程度分别为: CLIP-Zero 下降了 8.8 个百分点, CLIP-InfoNCE 下降了 10.16 个百分点; CLIP-SC 下降了 7.85 个百分点; CLIP-SC-HN 下降了 7.78 个百分点. 这进一步说明本文的模型具有更强的表示能力, 一定程度缓解了实体库扩大带来的负面影响.

4.2 训练曲线

为了进一步分析自对比学习、难负例对比学习对于模型训练的影响, 本文绘制了在 1.3 万实体库上训练时的 Accuracy 变化曲线, 如图 5 和图 6 所示. 图中的曲线为 5 个随机种子下的性能均值, 而颜色范围为误差区间.

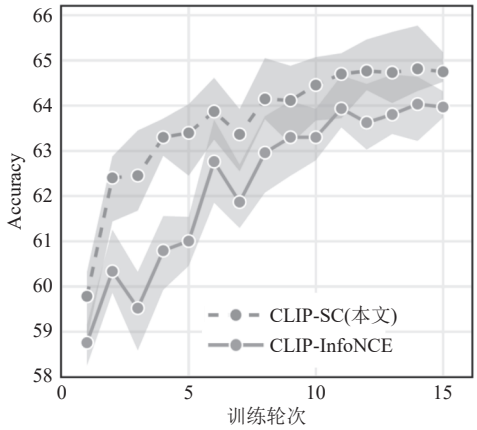


Fig. 5 Training curves after adding self-contrastive learning

图 5 添加自对比学习后的训练曲线

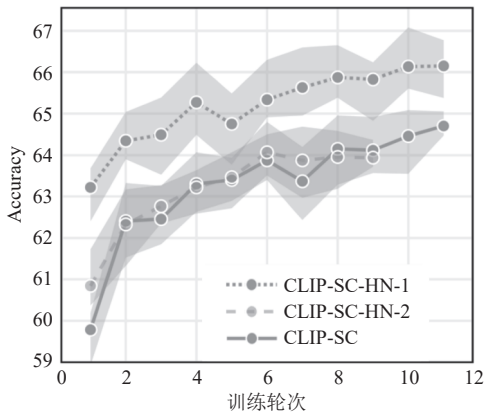


Fig. 6 Training curves after adding hard-negatives

图 6 添加难负例后的训练曲线

图 5 展示了添加自对比学习对训练过程的影响. 可以发现添加自对比学习 CLIP-SC 比只使用 InfoNCE 损失 CLIP-InfoNCE 在前期收敛更快, 且性能上下限以及均值都更高.

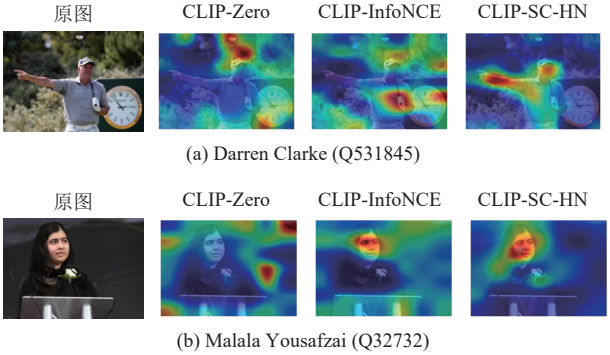
图 6 展示了添加难负例对训练过程的影响. 这里采用了 2 种模型初始化策略: 1)CLIP-SC-HN-1.用 CLIP-

SC初始化，相当于首先不添加难负例训练至收敛，之后添加难负例继续训练至收敛. 2)CLIP-SC-HN-2. 用CLIP-Zero初始化，相当于在训练开始时就添加难负例. 训练过程中都采用了早停策略，当2轮性能提升小于0.2%时停止训练. 可以发现，当采用CLIP-SC-HN-2策略，即一开始就加入难负例训练时，性能与不添加难负例的CLIP-SC没有明显差异. 但当采用CLIP-SC-HN-1策略，即先训练至收敛后再添加难负例，性能明显优于CLIP-SC.根据前人研究[43−44]，难负例的使用并不是随意的，直接不加限制地使用可能会导致性能下降. 根据文献[36]的理论推导，当温度系数较小时，梯度由难负例主导. 如果在训练早期就加入难负例，可能导致网络过于关注这些样本，反而无法应对大量简单样本. 因此，应该采用一种由易到难的训练策略，先只用简单样本训练至收敛，待模型具备一定判别能力后再添加难负例，进一步学习如何区分相似的困难样本.

4.3 可视化分析

本文进一步通过可视化分析，辅助证实了本文模型能提升对视觉实体的编码能力. 具体而言，本文使用了类别激活映射(class activation mapping, CAM)[45]技术，观察图像编码器对输入图像中各个区域的响应情况. CAM依据神经网络的输出权重，生成一张与输入图像大小相同的热力图，用来显示每个像素对模型预测结果的重要程度.

图7展示了在2个图像实体上，不同模型应用CAM后的可视化结果(CLIP-SC与CLIP-SC-HN的可视化效果没有明显差异，故未列出). 图7(a)的4张图像是实体Darren Clarke在打高尔夫球的照片，而图7(b)的4张图像是Malala Yousafzai在演讲的照片. 在4张图像中，第一张为原图，第二、三、四张依次为CLIP-Zero、CLIP-InfoNCE、CLIP-SC-HN的CAM热力图. 可以发现，CLIP-Zero主要关注在背景上，对人物不够关注. CLIP-InfoNCE部分关注了人物实体，但依然将很多注意力分散在背景上，而CLIP-SC-HN则更加明确地关注了人物实体，而不是背景. 通过以上可视化分析，可以得出结论：本文模型可以过滤掉无关的背景信息，重点关注在视觉实体本身上，从而学会区分实体的细粒度差异，提升链接性能.



(a) Darren Clarke (Q531845)

(b) Malala Yousafzai (Q32732)

Fig. 7 Class activation mapping of the image encoder

图7 图像编码器的类别激活映射图

5 结 论

本文提出了一种新的图文跨模态实体链接模型. 该模型在基于InfoNCE损失的对比学习任务基础上，在图像到文本的匹配中添加了图像的自对比学习任务，在文本到图像的匹配中添加了针对难负例图像的对比学习任务. 该模型给予了图像模态更多的关注，从而增强了图像实体向量的表示能力. 在WikiPerson数据集上的实验表明，本文模型明显提升了链接的性能. 进一步的可视化实验表明，本文模型在图像模态可以过滤掉无关的背景信息干扰，重点关注在目标实体本身上，从而改善了对视觉实体的表示能力.

未来的工作包括2个方面：1)将实体从人物扩展到其他非人物实体，例如交通工具、建筑物、动植物等；2)扩展到一张图像中包含多个视觉实体的情况.

作者贡献声明：王苑铮提出研究思路，设计实验，撰写论文；孙文祥实施实验，分析数据；范意兴、廖华明、郭嘉丰提出指导意见，修改论文.

参 考 文 献

[1] Sun Wen, Fan Yixing, Guo Jiafeng, et al. Visual named entity linking: A new dataset and a baseline [C] // Proc of the Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2022: 2403−2415

[2] Liu Fuxiao, Wang Yinghan, Wang Tianlu, et al. Visual news: Benchmark and challenges in news image captioning [C] // Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 6761−6771

[3] Abdelnabi S, Hasan R, Fritz M. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 14920−14929

[4] Bouquet P, Stoermer H, Giacomuzzi D, et al. OKKAM: Enabling a Web of entities [C/OL] // Proc of the WWW Workshop I3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web. New York: ACM, 2007 [2024-06-24]. https://ceur-ws.org/Vol-249/submission_150.pdf

[5] Milne D, Witten I. Learning to link with Wikipedia [C] // Proc of the 17th ACM Conf on Information and Knowledge Management. New York: ACM, 2008: 509−518

[6] Moro A, Raganato A, Navigli R. Entity linking meets word sense disambiguation: A unified approach[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 231−244

[7] Ferragina P, Scaiella U. Tagme: On the-fly annotation of short text fragments (by Wikipedia entities) [C] // Proc of the 19th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2010: 1625−1628

[8] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C] // Proc of the Annual Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 4171−4186

[9] Yamada I, Washio K, Shindo H, et al. Global entity disambiguation with BERT [C] // Proc of the Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2022: 3264−3271

[10] Cao N, Izacard G, Riedel S, et al. Autoregressive entity retrieval [C/OL] // Proc of the 9th Int Conf on Learning Representations. OpenReview. net, 2021 [2024-06-24]. https://openreview.net/forum?id=5k8F6UU39V

[11] Logeswaran L, Chang Mingwei, Lee K, et al. Zero-shot entity linking by reading entity descriptions [C] // Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 3449−3460

[12] Wu L, Petroni F, Josifoski M, et al. Scalable zero-shot entity linking with dense entity retrieval [C] // Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 6397−6407

[13] Moon S, Neves L, Carvalho V. Multimodal named entity disambiguation for noisy social media posts [C] // Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 2000−2008

[14] Zhou Xingchen, Wang Peng, Li Guozheng, et al. Weibo-mel, wikidata-mel and richpedia-mel: Multimodal entity linking benchmark datasets [C] // Proc of the 6th Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction. Berlin: Springer, 2021: 315−320

[15] Wang Sijia, Li A, Zhu Henghui, et al. Benchmarking diverse-modal entity linking with generative models [C] // Proc of the 61st Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 7841−7857

[16] Zhang Gongrui, Jiang Chenghuan, Guan Zhongheng, et al. Multimodal entity linking with mixed fusion mechanism [C] // Proc of the 28th Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2023: 607−622

[17] Wang Xuwu, Tian Junfeng, Gui Min, et al. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types [C] // Proc of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL 2022: 4785−4797

[18] Zhang Li, Li Zhixu, Yang Qiang. Attention-based multimodal entity linking with high-quality images [C] // Proc of the 26th Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2021: 533−548

[19] Yang Chengmei, He Bowei, Wu Yimeng, et al. MMEL: A joint learning framework for multi-mention entity linking [C] // Proc of the 39th Conf on Uncertainty in Artificial Intelligence. New York: PMLR, 2023: 2411−2421

[20] Wang Peng, Wu Jiangheng, Chen Xiaohang. Multimodal entity linking with gated hierarchical fusion and contrastive training [C] // Proc of the 45th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2022: 938−948

[21] Adjali O, Besançon R, Ferret O, et al. Building a multimodal entity linking dataset from tweets [C] // Proc of the 12th Language Resources and Evaluation Conf. Paris: European Language Resources Association, 2020: 4285−4292

[22] Adjali O, Besançon R, Ferret O, et al. Multimodal entity linking for tweets [C] // Proc of the 42nd European Conf on Information Retrieval. Berlin: Springer, 2020: 463−478

[23] Gan Jingru, Luo Jinchang, Wang Haiwei, et al. Multimodal entity linking: A new dataset and a baseline [C] // Proc of the 29th ACM Int Conf on Multimedia. New York: ACM, 2021: 993−1001

[24] Yang Zhishen, Okazaki N. Image caption generation for news articles [C] // Proc of the 28th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2020: 1941−1951

[25] Hu Anwen, Chen Shizhe, Jin Qin. Icecap: Information concentrated entity-aware image captioning [C] // Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 4217−4225

[26] Fukui A, Park D, Yang D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding [C] // Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 457−468

[27] Kim J, On K, Lim W, et al. Hadamard product for low-rank bilinear pooling [C/OL]// Proc of the 4th Int Conf on Learning Representations. OpenReview. net, 2016 [2024-06-24]. https://openreview.net/forum?id=r1rhWnZkg

[28] Younes H, Cadene R, Cord M, et al. Mutan: Multimodal tucker fusion for visual question answering [C] // Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 2631−2639

[29] Younes H, Cadene R, Thome N, et al. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection [C] // Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 8102−8109

[30] Szegedy C, Ioffe S, Vanhoucke V, et al. Inceptionv4, inception-ResNet and the impact of residual connections on learning [C] // Proc of the 31st AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 4278−4284

[31] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735−1780

[32] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Proc of the 31st Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2017: 6000–6010

[33] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [C/OL] // Proc of the 9th Int Conf on Learning Representations. OpenReview. net, 2021 [2024-06-24]. https://openreview.net/forum?id=YicbFdNTTy

[34] Radford A, Kim J, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] // Proc of the 38th Int Conf on Machine Learning. New York: PMLR, 2021: 8748−8763

[35] Liu Yinhan, Ott M, Goyal N, et al. RoBERTa: A robustly optimized bert pretraining approach [C/OL] // Proc of the 7th Int Conf on Learning Representations. OpenReview. net, 2019 [2024-06-24]. https://openreview.net/forum?id=SyxS0T4tvS

[36] Wang Feng, Liu Huaping. Understanding the behaviour of contrastive loss [C] // Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 2495−2504

[37] Chen Ting, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations [C] // Proc of the 37th Int Conf on Machine Learning. New York: PMLR, 2020: 1597−1607

[38] He Kaiming, Fan Haoqi, Wu Yuxin, et al. Momentum contrast for unsupervised visual representation learning [C] // Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 9729−9738

[39] Cao Qiong, Shen Li, Xie Weidi, et al. VGGFace2: A dataset for recognising faces across pose and age [C] // Proc of the 13th IEEE Int Conf on Automatic Face & Gesture Recognition. Piscataway, NJ: IEEE, 2018: 67−74

[40] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[41] Loshchilov I, Hutter F. Decoupled weight decay regularization [C/OL] // Proc of the 7th Int Conf on Learning Representations. OpenReview. net, 2019 [2024-06-24]. https://openreview.net/forum?id=Bkg6RiCqY7

[42] Jeff J, Matthijs D, Herve J. Billion-scale similarity search with GPUs [J]. IEEE Transactions on Big Data, 2019, 7(3): 535−547

[43] Robinson J, Chuang C, Sra S, et al. Contrastive learning with hard negative samples [C/OL] // Proc of the 9th Int Conf on Learning Representations. OpenReview. net, 2021 [2024-06-24]. https://openreview.net/forum?id=CR1XOQ0UTh-

[44] Kalantidis Y, Sariyildiz M, Pion N, et al. Hard negative mixing for contrastive learning [C] // Proc of the 34th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 21798−21809

[45] Zhou Bolei, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2921−2929



Wang Yuanzheng, born in 1997. PhD candidate. His main research interests include information retrieval and natural language processing.

王苑铮, 1997 年生. 博士研究生. 主要研究方向为信息检索、自然语言处理.



Sun Wenxiang, born in 1998. Master. His main research interests include information retrieval and natural language processing.

孙文祥, 1998 年生. 硕士. 主要研究方向为信息检索、自然语言处理.



Fan Yixing, born in 1990. Associate professor. His main research interests include information retrieval and natural language understanding.

范意兴, 1990 年生. 副研究员. 主要研究方向为信息检索、自然语言理解.



Liao Huaming, born in 1972. Associate professor. Her main research interests include big data application, information retrieval, and distributed data processing.

廖华明, 1972 年生. 副研究员. 主要研究方向为大数据应用、信息检索、分布式数据处理.



Guo Jiafeng, born in 1980. Professor. His main research interests include representation learning, and neural models for information retrieval and filtering.

郭嘉丰, 1980 年生. 研究员. 主要研究方向为表示学习、针对信息检索与过滤的神经模型.