潘超凡, 童骁, 彭焘,等. 基于对比学习和迁移学习的自动音频字幕系统[J]. 智能计算机与应用,2025,15(3):1-6. DOI:10.20169/j. issn. 2095-2163. 250301

基于对比学习和迁移学习的自动音频字幕系统

潘超凡[1], 童 骁[1], 彭 焘[1], 李圣辰[2], 朱晨阳[1], 邵 曦[1]
(1 南京邮电大学 通信与信息工程学院, 南京 210003; 2 西交利物浦大学 智能工程学院, 江苏 苏州 215123)

摘 要: 自动音频字幕是一项跨模态翻译任务,旨在使用自然语言来描述一段音频剪辑的内容。该任务近年来受到国内外广泛关注。现有的自动音频字幕系统通常基于编码器-解码器结构,而数据稀缺问题始终是自动音频字幕系统训练面临的一大难题。针对这一问题,文中提出一种新的模型架构,称为预编码器-编码器-解码器模型。在预编码器阶段,采用对比学习的方法从原始音频和配对文本数据中提取自监督信号,同时采用了迁移学习加快训练,并为编码器提供初始化参数。在 Clotho 数据集上的实验结果表明,文中提出的系统与基线系统相比性能显著提升。

关键词: 自动音频字幕; 跨模态翻译; 对比学习; 迁移学习; 音频剪辑
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)03-0001-06

Automated audio captioning system based on contrastive learning and transfer learning

PAN Chaofan[1], TONG Xiao[1], PENG Tao[1], LI Shengchen[2], ZHU Chenyang[1], SHAO Xi[1]
(1 School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; 2 School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China)

Abstract: Automated audio captioning is a cross-modal translation task that aims to use natural language to describe the content of an audio clip. This task has received widespread attention both domestically and abroad in recent years. Existing automated audio captioning systems typically rely on an encoder-decoder structure, with data scarcity being a major challenge for training such systems. To address this issue, the paper proposes a new model architecture called the pre-encoder-encoder-decoder model. In the pre-encoder stage, contrastive learning is used to extract self-supervised signals from raw audio and paired text data, while transfer learning is employed to accelerate training and provide initialization parameters for the encoder. Experimental results on the Clotho dataset show significant performance improvements of the proposed system compared to the baseline system.

Key words: automated audio captioning; cross-modal translation; contrastive learning; transfer learning; audio clip

0 引 言

自动音频字幕[1](Automated Audio Captioning, AAC)是一项涉及音频和文本的多模态任务,能将输入的音频信号转换为相应的自然语言描述、即字幕。相比于声音事件检测和声学场景分类研究,自动音频字幕描述了更广泛的信息,包括声音事件的识别、声学场景背景的识别以及对象和环境的概念和物理特性[2]。目前,音频字幕自动生成技术已经应用到生活中的许多方面,如实时音频文本转换,为电视中的声音配字幕,帮助听力受损的人员理解周围环境中的声音,分析智能城市中的声音以进行安全监控等。基于此,随着未来自动音频字幕研究的不断深入,应用场景将会更为丰富,人们的生活也将更加智能化。

2017 年,Drossos 等学者[1]正式提出了自动音频字幕任务。实验采用了基于门控循环单元(Gate Recurrent Unit,GRU)的编码器-解码器结构。由于

基金项目: 国家科技创新 2030—"新一代人工智能"重大项目(2020AAA0106200); 国家自然科学基金(61936005,62001038); 姑苏领军人才青年人才创新项目(ZXL2022472)。

作者简介: 潘超凡(1999—),男,硕士研究生,主要研究方向:自动音频字幕。
通信作者: 邵 曦(1976—),男,博士,教授,博士生导师,主要研究方向:音乐内容分析与检索,多媒体跨模态分析,多媒体跨平台个性化推荐等。Email:shaoxi@ njupt. edu. cn。

收稿日期: 2023-09-01

受制于数据集质量的原因,实际模型所产生的效果并不好,生成的字幕同原始字幕差距较大,但是以编码器-解码器为基础框架来进行音频字幕自动生成的探索,成为了后来绝大部分音频字幕工作的基础。

现有的自动音频字幕系统通常基于编码器-解码器体系结构[3-6]。音频数据首先被编码器编码成一种潜在的表示形式,同所配对的文本表示形式对齐,再通过解码器来生成标题。要想训练出一个良好的自动音频字幕系统,数据集的大小与质量尤为重要。

数据集 AudioCaps[7] 和 Clotho[8] 的相继问世,极大程度上提高了自动音频字幕模型生成字幕的质量,然而与自动图像字幕[9-11]的训练相比,在训练自动音频字幕系统时,数据稀缺的问题始终存在,而这可能导致不准确的特征表示以及音频文本特征的不匹配,从而生成同原始字幕相差较大甚至错误的字幕。

预训练音频神经网络 PANNs [12] (Pretrained Audio Neural Networks) 的出现则有效缓解这一问题。PANNs 是在大规模音频数据集 AudioSet 上训练得到的用于音频模式识别的神经网络,在下游任务中展现出强大的性能。近年来的一些自动音频字幕系统[13],通过迁移学习的方法,利用 PANNs 直接初始化编码器参数,模型性能明显优于从头开始训练的传统方法。因为 PANNs 主要针对单音频模态,而自动音频字幕任务属于多模态任务,通过 PANNs 提取的音频特征在进入解码端前同文本模态特征没有建立任何联系,所以在传统的编码器-解码器架构中,直接进入解码端的的音频嵌入特征可能不是适用于音频字幕这一多模态任务的最佳表示。

为了不过度依赖外部数据集和外部预训练模型,本文在原有迁移学习的部分基础上,引入了对比学习来提高对原始数据集的利用。对比学习是一种自我监督的范式[14-15],可以帮助模型获得高质量的表示。受计算机视觉[16-17](Computer Vision,CV)和自然语言处理[18-19](Natural Language Processing,NLP)中一些工作的启发,本文提出了一种新的模型架构,称为预编码器-编码器-解码器模型。在预编码器中,通过对比样本,利用从原始音频-文本配对数据中得到的自监督信号,获取音频和文本之间的对应关系。更准确地说,是将原始匹配的音频-文本对作为正样本,构造不匹配的音频-文本对作为负样本。然后,设计了一个对比学习目标,旨在最大化正样本与负样本之间表示的差异。通过这种方式,在有限的数据量训练的情况下,可以进一步提高潜在表示质量和音频与文本之间的对齐。最后通过迁移学习,将预编码器中训练好的音频编码器迁移到编码器中,并针对自动音频字幕任务进行微调。

1 模型系统介绍

本文以编码器-解码器为基础,额外增加了预编码器模块。预编码器采用了一种双编码器模型,结构如图 1 所示。采用预训练好的 10 层卷积神经网络[20](Convolutional Neural Networks,CNN)作为音频编码器,预训练好的 Word2Vec 模型[21]作为文本编码器。编码器-解码器中编码器同样使用 10 层的卷积神经网络,区别在于编码器是通过预编码器中的音频编码器得到的。解码器使用完整 Transformer[22]中的解码器部分。整个模型架构如图 2 所示。

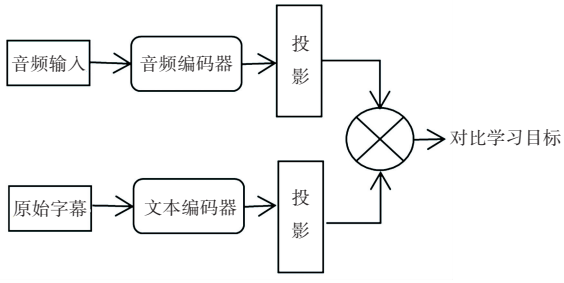

图 1 预编码器结构

Fig. 1 Architecture of the pre-encoder

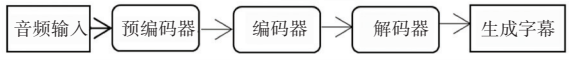

图 2 模型架构

Fig. 2 Architecture of the model

1.1 预编码器

预编码器由音频编码器和文本编码器组成,卷积神经网络现已广泛应用于音频处理相关的工作中,并在提取音频特征方面显示出强大能力。音频编码器采用 10 层的卷积神经网络提取音频信号的特征,该网络由 4 个卷积块组成。每个卷积块由 2 组卷积核大小为 3×3 的卷积层和 1 组核大小为 2×2 的池化层组成,为加快训练速度以及提高模型精度,每个卷积层后使用批归一化[23](Batch Normalization,BN)并使用 *ReLU* 作为激活函数。卷积块的通道数依次为 64,128,256,512。文本编码器采用 Word2Vec,将文本描述转换为单词嵌入序列。简单起见,研究中采用了一个预先训练过的 Word2Vec,是对 Clotho 数据集中所有字幕中的单词

进行训练的。利用模型的隐层权重，可以获得每个单词的词向量表达，在向量空间中，具有相似语义或语法结构的单词距离较近，因而可以协助解码器生成更为自然流畅的语句。

1.2　编码器

在过往的研究中，卷积神经网络（CNN）展现出良好的音频特征提取能力。为了提取更多、更复杂的特征且有效防止过拟合，选用 10 层的卷积神经网络 CNN_{10} 作为编码器，具体的结构组成同上述预编码器中音频编码器部分。

另一方面，考虑到 PANNs 在提取不同下游音频识别任务的音频信号的潜在表示方面展现出的强大能力，在传统的编码器-解码器中，直接采用 PANNs 中对应的 CNN_{10} 模型来初始化编码器参数。而在预编码器-编码器-解码器中，PANNs 主要用于初始化预编码器中的音频编码器模型参数，完成对比学习后的最佳预编码器中的音频编码器模型参数通过迁移学习来初始化编码器模型参数。

1.3　解码器

Transformer 模型在自然语言处理的一系列任务中表现出色，为了将音频特征和词嵌入连接起来，研究使用 Transformer 作为解码器。一个标准的 Transformer 模型由编码端和解码端组成，设计中主要将用于处理文本序列，而不是时间序列。因此，研究中仅使用带有交叉注意力的解码端来生成字幕文本。

具体而言，解码器由 2 部分组成，分别为词嵌入层和标准的 Transformer 解码端。其中，Transformer 解码端共计 4 层，每层有 4 个注意力头，最后使用一个线性层通过 Softmax 参照词汇表输出概率分布。

1.4　对比学习

音频字幕系统训练始终面临着很难在网络上获取大量高质量的音频-字幕对的问题，早期有学者尝试使用一些数据增强的手段来提高对原始数据集的利用[24-25]。然而，增强后的数据可能会丢失原始数据的某些重要特征，同时可能会引入一些噪声或错误信息，干扰模型的学习过程，无法达到对原始数据集的充分利用。因此，研究尝试在预编码器中使用对比学习的方法，利用同一批中其他样本来构建不匹配的音频-字幕对并作为负样本，通过对比学习来提高负样本所对应的音频模态和文本模态表示的差异，从而提高音频-文本表示的质量以及音频和文本之间的对齐性。

为此分析推得，在批大小为 N 的音频-字幕对中，每一批样本共包含 N 组音频-字幕对 (A_1, C_1)，(A_2, C_2)，(A_3, C_3)，…，(A_N, C_N)，预编码器首先将音频和字幕分别进行编码嵌入：

$$a_i = CNN_{10}(A_i) \tag{1}$$

$$c_i = Word2Vec(C_i) \tag{2}$$

其中，a 表示音频嵌入；c 表示文本嵌入；下标 i 表示第 i 组；具有相同下标的 a 和 c 表示来自同一组音频 - 字幕对。

然后，分别将音频编码嵌入和字幕编码嵌入映射到相同维度，并通过全连接层统一目标维度，再通过 $L2$ 正则化防止过拟合：

$$a_i^F = FC(a_i) \tag{3}$$

$$c_i^F = FC(c_i) \tag{4}$$

$$a_i^P = L2(a_i^F) \tag{5}$$

$$c_i^P = L2(c_i^F) \tag{6}$$

将音频样本和文本样本编码投影后的向量之间的点积作为相似度度量，同一组音频-字幕对看作正样本，不同组音频 - 字幕对看作负样本，以 *InfoNCE* 损失为训练损失，为了更充分地利用原始数据集，分别从音频和文本两个角度来构造正负样本，具体计算如下：

$$s_{i,j} = a_i^P \cdot c_j^P \tag{7}$$

$$L_{ai} = -\log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{N} \exp(s_{i,j}/\tau)} \tag{8}$$

$$L_{cj} = -\log \frac{\exp(s_{j,j}/\tau)}{\sum_{i=1}^{N} \exp(s_{i,j}/\tau)} \tag{9}$$

$$L = \frac{1}{N}\left(\sum_{i=1}^{N} L_{ai} + \sum_{j=1}^{N} L_{cj}\right) \tag{10}$$

其中，τ 表示一个温度参数，用于调整相似度值的尺度；$s_{i,j}$ 表示计算第 i 组音频样本与第 j 组字幕样本之间的相似度；L_{ai} 是从音频模态角度出发，比较每一组中的音频样本和所有组中的字幕样本；L_{cj} 是从文本模态角度出发，比较每一组中的字幕样本和所有组中的音频样本。L 越小，表示正、负样本的特征表示间区分度越大，从 L_{ai} 和 L_{cj} 两个角度出发，可以提供更多的负样本，并有效提高正样本中音频和文本表示的对齐性。

1.5　迁移学习

迁移学习是一种机器学习技术，可以将已经训练好的模型应用到新的任务中[26]。该方法可以有效地利用现有数据和模型的知识来改善新任务的性

能。在传统的机器学习中,每个任务都需要从头开始训练一个新的模型,然而在实践中,却往往面临着数据量不足、时间成本高昂等问题。通过迁移学习,可以利用此前训练好的模型,加快新任务的训练,同时减少对大量标注数据的依赖。迁移学习主要用于涉及单一模态的任务,对于自动音频字幕这个跨模态(即音频到文本)的翻译任务,迁移学习的相关研究主要集中在音频模态上[6]。

本文主要使用 2 次迁移学习方法。第一次迁移学习方法是直接在预编码器中分别使用 PANNs 和预训练好的 Word2Vec 模型来初始化音频编码器和文本编码器模型参数。第二次迁移学习方法是将预编码器中训练好的音频编码器参数迁移到编码器中去。

2 实验与设置

2.1 数据集

研究中使用 DCASE2022 挑战赛的任务 6a 提供的数据集 Clotho (V2) 作为实验数据集。Clotho 数据集由持续时间为 15~30 s 的音频片段组成,每个音频片段对应 5 个由 8~20 个单词组成的字幕。Clotho (V2) 共有 6 972 个音频样本,对应 34 860 个字幕,根据构成字幕的单词集合划分为 4 个部分:3 839 个音频片段组成的开发集,1 045 个音频片段组成的验证集,1 045 个音频片段组成的评估集,1 043 个音频片段组成的测试集。研究中还合并了开发集和验证集,形成了由 4 884 个音频片段组成的新训练集,并在评估集上进行评估。

2.2 数据预处理

数据集中原始音频信号的采样率为 44.1 kHz,使用 1 024 个点、50%重叠的汉明窗口和 64 个梅尔滤波器组提取出 log-Mel 谱作为音频输入特征。为了统一编码维度,研究使用 0 来填充音频频谱到最大时间序列长度。数据集中的所有字幕都被转换为小写并删除标点符号。此外,又使用 2 个特殊的标记“<sos>”和“<eos>”分别作为起始标记和结束标记来填充标题。

2.3 模型训练

模型训练主要分为 2 步,分别为预编码器训练和编码器-解码器训练。

在预编码器训练中,以 *InfoNCE* 损失为目标函数[27],使用 Adam 优化器,学习率为 0.001,共训练 80 个周期,批处理大小为 32。使用学习率调度器 ReduceLROnPlateau 来调整学习率,如果连续 5 个周期监测目标仍未能改善,学习率降低为原来的 0.1,最小学习率设置为 0.000 001。

在编码器-解码器训练中,以交叉损失熵为目标函数,使用 AdamW 优化器,在前 5 个周期使用热身训练,学习率从 0 线性增长到 0.001,此后每 5 个周期下降为原来的 0.1,共训练 30 个周期,批处理大小为 16。为了缓解过拟合问题,该部分额外应用了大小为 0.2 的随机丢失 *dropout*。

2.4 评估指标

实验中根据上述基于对比学习和迁移学习的自动音频字幕模型生成的字幕质量好坏来评估模型的性能。概括来讲,是将模型生成字幕与初始字幕进行对比,通过 $BLEU_n$、*ROUGE*、*METEOR*、*CIDEr*、*SPICE* 和 *SPIDEr* 这 6 项指标来计算得分,根据得分对模型的性能做出全面的评估。具体介绍如下。

(1) $BLEU_n$[28]:主要思想在于利用 n-gram(文本序列中连续的 n 个单词)技术对模型生成语句和参考语句进行比较,通过计算相互之间的重合度来评估语句的质量。具体而言,当 n 较小时,$BLEU_n$ 可反映模型生成的单词和短语的准确性;而当 n 较大时,则能够在一定程度上衡量句子的流畅性。通常情况下,常见的 n 取值范围为 1~4。

(2) *ROUGE*[29]:主要是用于自动生成文摘和机器翻译等任务的评估,其着重关注的是文本内容的重合度。与 $BLEU_n$ 不同,*ROUGE* 只考虑召回率而不涉及准确率的计算。

(3) *METEOR*[30]:是一种广泛用于自然语言翻译质量评估的指标,主要考虑了单词层次和句子层次上的匹配问题,并且通过使用外部数据拓展自身的词库,可以考虑同义词、词性等更加丰富的语言学信息。与 $BLEU_n$ 和 *ROUGE* 不同,*METEOR* 还考虑了参考译文中的语序问题,这意味着能进一步处理翻译结果与参考译文之间的单词顺序不同的情况。最终,*METEOR* 同时计算翻译结果的准确率和召回率,并通过加权平均得到最终的评估分数。

(4) *CIDEr*[31]:是一种用于评估自然语言生成模型的指标,旨在比较自动生成的字幕与原始字幕间的相似度。通过计算生成字幕与原始字幕之间的 n-gram 重叠程度来衡量自动生成的字幕的质量。该指标使用了多个原始字幕并进行加权处理,可用来反映模型输出与参考描述之间的一致性。

(5) *SPICE*[32]:尽管 *SPICE* 最初是针对图像字幕生成模型的评价指标,但仍可用于音频字幕生成模型的质量评估。在音频字幕任务中,*SPICE* 可将

自动生成的字幕转化为场景图，并基于语义准确性和完整性对场景图与原始音频之间的关联得分进行计算。通过使用 *SPICE* 对音频字幕生成模型进行评估，可以了解自动生成字幕与原始音频之间的相匹配程度，从而衡量该模型的质量。

（6）*SPIDEr*[33]：是 *CIDEr* 指标和 *SPICE* 指标的加权平均，综合了 *CIDEr* 和 *SPICE* 两个指标的优点。*CIDEr* 主要关注生成字幕的流畅性和一致性，而 *SPICE* 则主要考虑生成字幕是否与原始音频片段内容相符。通过将这 2 个指标的加权平均，*SPIDEr* 可以对音频字幕模型进行全面的质量评估。

2.5　实验结果分析

研究中选用 2022DCASE 官方提供的基线模型为对照，基线模型使用一个完整的 Transformer。这是一个标准的序列到序列 Transformer，具有 6 个编码器层和 6 个解码器层。通过迁移学习，编码器从维度为 128 的预训练 VGGish 模型中获取输入嵌入，再由仿射层转换为维度为 768 的嵌入，并输出与输入序列长度相同的嵌入序列。编码器的每一层针对输入表示的每个时间步输出 768 个特征。为此，共设计了 2 组实验。第一组实验是在传统编码器-解码器架构基础上搭建轻量级 CNN-Transformer 音频字幕系统，同时使用迁移学习，利用 PANNs 初始化编码器模型参数。第二组实验是在第一组实验基础上增加预编码器模块，使用对比学习结合迁移学习优化预编码器中的音频编码器，再迁移到编码器中针对自动音频字幕表示进行微调。

表 1 统计了预编码器所用参数以及编码器-解码器架构下基线模型和实验 1 中模型使用的参数量，表 2 统计了基线模型以及实验 1 和实验 2 在 Clotho V2 评估集上各项指标的得分。比较基线模型和实验 1 模型的参数量和各项指标得分，仿真实验所用的 CNN-Transformer 模型规模远小于基线完整 Transformer 的模型规模，但在性能上明显优于基线模型。同实验 1 相比，实验 2 在各项指标的得分上有所提升，这也进一步证明了引入预编码器并应用对比学习策略的有效性。

表 1　模型参数统计

Table 1　Model parameter statistics

模型	参数/M
预编码器	6
CNN-Transformer	9
标准 Transformer	140

表 2　实验结果

Table 2　Experimental results

模型	基线模型	实验 1	实验 2
$BLEU_1$	0.555	0.564	0.613
$BLEU_2$	0.358	0.362	0.402
$BLEU_3$	0.239	0.240	0.273
$BLEU_4$	0.156	0.156	0.175
ROUGE	0.364	0.372	0.377
METEOR	0.164	0.170	0.172
CIDEr	0.358	0.384	0.412
SPICE	0.109	0.117	0.121
SPIDEr	0.233	0.251	0.266

3　结束语

本文针对自动音频字幕系统训练的数据稀缺问题，提出了预编码器-编码器-解码器结构框架，以更好地学习跨模态表示。通过对比学习的方法，在预编码器中进行自监督学习，可以有效地利用多模态数据中的信息，从而提取出更有意义和鲁棒性的特征表示。再通过迁移学习的方式，将训练后的预编码器的音频编码器模型参数迁移到编码器中去。实验结果表明，和传统的编码器-解码器结构相比，经过预编码器迁移学习的编码器提取的音频特征同原始的文本特征更加匹配，有效提高了模型各项指标的得分。在未来的研究工作中，将会探讨更有效的对比学习方法，扩充更多的数据样本来学习更好的音频-文本表示。

参考文献

[1] DROSSOS K, ADAVANNE S, VIRTANEN T. Automated audio captioning with recurrent neural networks [C]//2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). Piscataway, NJ: IEEE, 2017: 374-378.

[2] NGUYEN K, DROSSOS K, VIRTANEN T. Temporal sub-sampling of audio feature sequences for automated audio captioning [J]. arXiv preprint arXiv, 2007.02676, 2020.

[3] KOIZUMI Y, MASUMURA R, NISHIDA K, et al. A transformer-based audio captioning model with keyword estimation [J]. arXiv preprint arXiv, 2007.00222, 2020.

[4] WU Mengyue, DINKEL H, YU Kai. Audio caption: Listen and tell [C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019). Piscataway, NJ: IEEE, 2019: 830-834.

[5] TRAN A, DROOSOS K, VIRTANEN T. WaveTransformer: A novel architecture for audio captioning based on learning temporal and time-frequency information [J]. arXiv preprint arXiv, 2010.11098, 2020.

[6] MEI Xinhao, HUANG Qiushi, LIU Xubo, et al. An encoder-

decoder based audio captioning system with transfer and reinforcement learning[J]. arXiv preprint arXiv, 2108.02752, 2021.
[7] KIM C D, KIM B, LEE H, et al. Audiocaps: Generating captions for audios in the wild[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). ACL,2019: 119-132.
[8] DROSSOS K, LIPPING S, VIRTANEN T. Clotho: An audio captioning dataset[C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway,NJ: IEEE, 2020: 736-740.
[9] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//Proceedings of 13th European Conference on Computer Vision(ECCV 2014). Cham: Springer, 2014: 740-755.
[10]宋光慧. 基于迁移学习与深度卷积特征的图像标注方法研究[D]. 杭州:浙江大学,2017.
[11]刘培. 基于深度学习的图像字幕生成研究[D]. 成都:四川大学,2021.
[12]KONG Qiuqiang, CAO Yin, IQBAL T, et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2880-2894.
[13]陈耕耘,李圣辰,邵曦,等. 基于迁移学习与强化学习的自动音频标注系统[J]. 复旦学报(自然科学版),2022,61(5):520-526.
[14]韩滕跃,牛少彰,张文. 基于对比学习的多模态序列推荐算法[J]. 计算机应用,2022,42(6):1683-1688.
[15]卢绍帅,陈龙,卢光跃,等. 面向小样本情感分类任务的弱监督对比学习框架[J]. 计算机研究与发展,2022,59(9):2003-2014.
[16]CHEN Ting, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[J]. arXiv preprint arXiv, 2002.05709,2020.
[17]孙浩,徐延杰,陈进,等. 基于自监督对比学习的深度神经网络对抗鲁棒性提升[J]. 信号处理,2021,37(6):903-911.
[18]HUANG Qiushi, KO T, TANG H L, et al. Token-level supervised contrastive learning for punctuation restoration[J]. arXiv preprint arXiv,2107.09099, 2021.
[19]徐守坤,徐坚,李宁,等. 基于 Sentence-Rank 的图像句子标注[J]. 计算机工程与应用,2019,55(2):121-127.
[20]CHEN Kun, WU Yusong, WANG Ziyue, et al. Audio captioning based on Transformer and pre-trained CNN[C]// Detection and Classification of Acoustic Scenes and Events 2020 Workshop. Tokyo, Japan: HITACHI,2020: 21-25.
[21]CHURCH K W. Word2Vec[J]. Natural Language Engineering, 2017, 23(1): 155-162.
[22]VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. Long Beach, USA: NIPS Foundation, 2017: 5998-6008.
[23]刘建伟,赵会丹,罗雄麟,等. 深度学习批归一化及其相关算法研究进展[J]. 自动化学报,2020,46(6):1090-1120.
[24]高友文,周本君,胡晓飞. 基于数据增强的卷积神经网络图像识别研究[J]. 计算机技术与发展,2018,28(8):62-65.
[25]陈文兵,管正雄,陈允杰. 基于条件生成式对抗网络的数据增强方法[J]. 计算机应用,2018,38(11):3305-3311.
[26]刘鑫鹏,栾悉道,谢毓湘,等. 迁移学习研究和算法综述[J]. 长沙大学学报,2018,32(5):28-31.
[27]OORD A, LI Yaze, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv, 1807.03748, 2018.
[28]汪菊琴,高俊涛. 基于实例的 BLEU 翻译评价方法[J]. 电脑知识与技术,2009,5(32):9035-9036.
[29]于俊婷,何宏业,刘伍颖,等. ROUGE-SN:基于跨越 N 元语法的机器翻译评测方法[J]. 数码设计,2017,6(3):1-5.
[30]BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments [C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. ACL,2005: 65-72.
[31]VEDANTAM R, LAWRENCE Z C, PARIKH D. Cider: Consensus-based image description evaluation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2015: 4566-4575.
[32]ANDERSON P, FERNANDO B, JOHNSON M, et al. Spice: Semantic propositional image caption evaluation[C]//Proceedings of 14th European Conference on Computer Vision(ECCV 2016). Cham: Springer, 2016: 382-398.
[33]LIU Siqi, ZHU Zhenhai, YE Ning, et al. Improved image captioning via policy gradient optimization of spider[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway,NJ:IEEE,2017: 873-881.