西北工業大學

本科毕业设计（论文）

题 目 ____基于深度学习的跨模态图像文本匹配研究____

专业名称_____人工智能_____

学生姓名_____牛远卓_____

指导教师_____牛凯_____

毕业时间_____2023.07.01_____

摘 要

随着互联网的普及和计算机视觉技术的发展，多模态数据大量产生。为了从丰富的多模态数据中挖掘信息，跨模态图像文本匹配受到了广泛关注。跨模态图像文本匹配是指通过比较和匹配图像和文本之间的语义相似性来建立它们之间的关联。该任务旨在将视觉模态和语句模态在同一空间中进行对齐，从而实现图像与文本之间的语义理解和相互关联。传统的跨模态图像文本匹配方法依赖手工设计的特征提取和匹配方法，需要数据特征的数量丰富且质量过硬，难以在更加广泛的应用场景落地。针对以上的问题，本项目实现了一种基于最大违规的深度模型图像文本匹配算法，此方法从损失函数的角度解决损失函数收敛慢的问题，以实现模型性能的提高。

本项目首先梳理了常见的图像文本匹配算法，并分析了现有方法存在的不足之处。其次，本项目详细介绍了深度学习技术在图像与文本匹配中的应用，包括深度学习模型的结构设计、训练技巧和性能评估方法。最后，本项目还详细介绍了基于深度学习的跨模态匹配方法的具体实现步骤。

为了验证所述方法的实际性能，本项目选取大规模的公开图像文本匹配数据集 Flickr30k 进行了多角度的实验。实验结果表明，本项目实现的基于最大违规的深度学习跨模态匹配方法具有很好的性能，比传统的基于平均违规的图像文本匹配方法在跨模态匹配上有更高的匹配精度和更好的效果。

关键词： 深度学习，图像与文本匹配，跨模态对齐，最大违规

ABSTRACT

With the popularity of the Internet and the development of computer vision technology, multimodal data are generated in large quantities. In order to mine information from the rich multimodal data, cross-membrane image text matching has received widespread attention. Cross-model image text matching is to encode the samples of the two modalities, image and text, separately to obtain their semantic representations, and also to have corresponding similarity calculation methods to calculate the similarity between these semantic representations. Traditional cross-modal image text matching methods rely on hand-designed feature extraction and matching methods, leading to excessive requirements on the quantity and quality of data features, making it difficult to implement in a wider range of application scenarios. To address the above issues, this thesis implements a maximum violation-based depth model image text matching algorithm. This method addresses the problem of slow convergence of the loss function from the perspective of the loss function in order to achieve improved model performance.

This thesis firstly compares the common image text matching algorithms and analyses the shortcomings of existing methods. Secondly, this thesis details the application of deep learning techniques in image and text matching, including the structural design of deep learning models, training techniques and performance evaluation methods. Finally, this thesis also details the concrete implementation steps of the cross-modal matching method based on deep learning.

In order to verify the practical performance of the described method, this thesis selects the large-scale image-text matching dataset Flickr30k to conduct experiments from multiple perspectives. The experimental results show that the maximum violation-based deep learning cross-modal matching method implemented in this thesis has good performance, with higher matching accuracy and better results on cross-modal matching than the traditional mean violation-based image text matching method.

KEY WORDS： deep learning, image-text matching, cross modal alignment, max violation

目录

第一章 绪论

1.1 研究背景

随着互联网的普及和计算机视觉技术的发展，图像和文本数据在人类社会中越来越广泛地应用。然而，图像和文本数据之间存在着固有的模态差异，因此如何建立起图像和文本之间的语义联系并实现它们之间的有效匹配，是计算机科学领域中一个被广泛研究的问题。

图像文本匹配是图像处理和自然语言处理的交叉领域，它是一种用来将图像和文本建立联系的技术。例如在电子商务平台上，需要对商品图片和商品描述进行匹配，以提供更好的搜索服务和购买体验；在社交媒体平台上，需要对照片和文本进行匹配，以便用户更好地分享和传播信息。

传统的图像文本匹配方法通常是基于手工设计的特征提取和匹配方法，这种方法往往受制于特征的质量和数量，无法充分挖掘数据的潜在特征。因此，基于深度学习的图像文本匹配方法应运而生，它可以自动地从数据中学习高质量的特征表示，并且在匹配任务中具有更好的表现。

近年来，基于深度学习的图像文本匹配方法已经取得了一系列重要的研究成果。然而，目前仍存在许多挑战和问题，例如如何进一步提高匹配的准确率和效率，如何解决跨模态匹配中存在的语义鸿沟等。

因此，本研究旨在使用深度学习方法，探究跨模态图像文本匹配的实现方式，提高匹配的准确率和效率，促进在电子商务、社交媒体等领域中的应用，为推动人工智能的发展做出贡献。

1.2 研究意义

近年来，随着人工智能技术的飞速发展，图像和文本信息的处理能力得到了极大的提高。而图像文本匹配作为一种用于自然语言处理和计算机视觉领域的任务，可以帮助人们更好地理解和处理多模态数据。因此，在本项目中，本实验将探索基于深度学习的跨模态图像文本匹配方法，以应对跨模态数据处理的需求。

首先，图像文本匹配在实际应用场景中具有很高的实用价值。例如，对于电商平台而言，可以根据用户输入的商品描述信息，通过图像文本匹配技术准确定位用户所需的商品，提高用户购物体验和销售额。在智能家居领域，通过图像识别和语音识别技术实现家居设备控制，也需要利用图像文本匹配技术实现普适性的控制操作。

其次，本研究所实现的跨模态图像文本匹配算法具有重要的理论意义。深度学习算法已经成为目前处理大规模数据的主流方法，同时深度学习对于跨模态信息处理也有着很好的适用性。因此本项目所探索的跨模态图像文本匹配算法，可为深度学习算法在跨模态信息处理领域的应用提供新的范例和思路，并为相关学科的理论研究提供一定的参考和借鉴。

最后，本项目的研究也有一定的推广应用价值。跨模态图像文本匹配算法在很多领域都有广泛的应用，如社交媒体分析、文化遗产保护、智能物联网等，本项目所实现的方法也有很大的推广价值，可以为相关领域的应用提供有力的支撑和参考。

因此，本研究通过基于深度学习的跨模态图像文本匹配方法的探索，旨在提升跨模态信息处理和深度学习算法的应用水平，同时为跨模态图像文本匹配的应用提供新的思路和方法，为相关领域带来新的理论和实用价值。

1.3 国内外研究现状

图像文本匹配技术是计算机视觉、自然语言处理和人工智能等领域交叉的研究热点之一。过去几年，国内外学者进行了广泛而深入的研究，取得了较为理想的成果。

在图像方面，传统的视觉识别技术主要以局部特征为基础，例如 SIFT[1]、SURF[2]和 HOG[3]等。然而，由于这些方法只关注图像的低-中级特征，其表征能力较弱，难以捕捉图像的高层语义信息。近年来深度学习技术在图像领域的应用，已经成为图像特征提取的主流方法。基于深度学习的卷积神经网络(CNN)方法可以自动学习图像的高层次特征，具有更强的表征能力。同时，GNN(图神经网络)等图像表示学习方法也获得了广泛的关注。这种新的视觉识别方法，能够从图像中学习出更丰富的高层次、语义化的特征，更好地理解图像内容。

在文本方面，传统的文本表示方法包括 Bag of Words(BoW)[4]和 TF-IDF[5]等方法。这些方法是基于特征选择和特征权重计算的思想，将文本转化成向量表示。但这种方法无法考虑语义信息，不能表示词汇关系。近年来，基于深度学习的文本表示方法受到了广泛的关注。Word2Vec[6]、GloVe[7]和 BERT[8]等方法利用词向量的方式将词汇转换为连续的向量，更好地表示了词汇之间的语义关系。

针对图像文本匹配问题，已有的研究主要分为两类：基于浅层特征的传统机器学习方法和基于深度学习的方法。基于浅层特征的方法，包括 SIFT[1]、SURF[2]和 HOG[3]等视觉特征和 TF-IDF[5]、LDA[9]等文本特征方法。这些方法主要关注图像的低级视觉特征和单纯字面的文本特征，限制了匹配精度和鲁棒性。基于深度学习的方法，使用深度网络来提取图像和文本的高级语义特征，该方法已经广泛地应用于图像文本匹配领域中。比如使用卷积神经网络(CNN)和递归神经网络(RNN)将图像和文本 embedding 到低维空间，再使用损失函数来计算匹配得分。

此外，还有一些基于注意力机制(Attention Mechanism)和交互注意力(Interactive Attention)的图像文本匹配算法，其主要思想是通过对图像和文本中的注意力区域进行动态更新，以实现更好的跨模态匹配。

最近，越来越多的基于深度学习的图像文本匹配算法被提出来，并在各种数据集上展示了出色的性能，如 Flickr30K[10]、MSCOCO[11]等。然而，跨模态匹配仍然面临着许多困难和挑战，如模型泛化能力差、耗费大量计算资源等问题。

总的来说，图像文本匹配的研究和应用领域正在快速发展，未来将有更多的基于深度学习的图像文本匹配方法被提出，为跨模态匹配问题提供更有效、更鲁棒的解决方案。

1.4 研究内容和目标

1.4.1 研究内容

本研究的主要内容是基于深度学习的跨模态图像文本匹配。跨模态图像文本匹配是一种重要的多媒体信息检索技术，涉及到图像和文本两种不同的形式，且以不同的方式表示相同的信息。因此，跨模态图像文本匹配涉及到图像和文本的特征提取、匹配算法及其评价指标等多方面的技术。

1.4.2 研究目标

本研究的主要目标是实现一种基于最大违规的深度学习跨模态图像文本匹配算法，实现图像和文本之间的匹配，以提高多媒体信息检索的准确性和效率。具体研究内容如下：

首先，介绍深度学习的基本理论和模型，包括深度神经网络、卷积神经网络和循环神经网络等。深度神经网络是深度学习的核心部分。它是由多个神经网络层组成，每个层都包含一些基本单元，例如卷积、池化、全连接等。卷积神经网络是深度学习中应用最广泛的模型之一，它通常用于图像分类、目标检测、分割等任务。CNN 的核心思想是通过探测局部特征，从而构建全局特征表达。它通过卷积层、池化层和全连接层等基本单元组成。

其次，介绍跨模态图像文本匹配的原理和方法，包括图像和文本的特征提取、匹配方法和评价指标等。在跨模态图像文本匹配中，特征提取是关键的一步。因为特征的质量直接影响到匹配准确性。本项目在 3.1 节详细介绍图像和文本的特征提取。图像文本匹配方法是一种将图像和文本信息进行关联的技术手段，旨在通过计算机算法实现对图像与文本的自动关联。本项目在 3.2 节详细介绍图像和文本的特征提取。评价指标应该能够反映算法的性能，便于比较不同算法之间的差异。本项目在 3.3 节详细介绍图像和文本的评价指标。

再次，设计并构建符合实际应用需要的跨模态图像文本匹配数据集，为后续算法的研究提供基础。本项目研究了 Flickr30k[10]和 MSCOCO[11]，并且通过对比选择前者构建数据集。本项目在 4.1 节详细介绍研究和使用的数据集。

然后，设计一种基于深度学习的跨模态图像文本匹配算法，包括模型架构设计、损失函数研究、训练与优化等。本项目实现了一种基于最大违规的深度学习跨模态图像文本匹配模型，该模型主要分为三部分：图像编码器、文本编码器和匹配层。本项目在 5.1 节详细介绍模型架构设计。本项目调研了两种常用的损失函数，分别是对比损失函数和三元组损失函数。本项目在 5.2 节详细介绍损失函数研究。本项目在 5.3 节详细介绍训练与优化。

接着，通过在构建的跨模态图像文本匹配数据集上进行实验，评估所实现的算法的准确性和效率，并进行结果分析和可视化展示。本项目在 6.2 节详细分析实验结果，并且在 6.3 节进行可视化展示。

最后，解释所实现的算法背后的原理和效果背后的原因，获得其背后的哲学原理。这也为之后的研究提供更多理论基础。本项目在 7.1 节详细介绍该部分内容。

通过以上研究内容和目标，本研究旨在实现一种基于深度学习的跨模态图像文本匹配算法，为多媒体信息检索技术的发展提供新的思路和方法，同时也为图像和文本的跨模态匹配领域的研究提供一定的参考价值。

1.5 研究流程与数据集

本研究主要采用深度学习技术进行图像文本匹配研究，涉及到以下方面的方法和数据集。

1.5.1 研究流程

首先，进行图像预处理。数据预处理步骤包括图像和文本的预处理。对于图像预处理，本研究采用了图像色彩空间转换和尺度归一化来使图像数据具有可比性。对于文本预处理，本研究采用了停用词过滤和文本清洗技术来去除无关和冗余信息，从而提高文本处理的效率和准确性。

其次，进行特征提取。特征提取是图像文本匹配的基础，本研究采用了基于深度学习的特征提取方法。对于图像特征提取，本研究采用了卷积神经网络(CNN)，并通过对预训练模型的微调来提高特征提取的准确性。对于文本特征提取，本研究采用了循环神经网络(RNN)和长短时记忆网络(LSTM[20])来提取文本的语义特征。

再次，进行模型匹配。本研究采用了基于深度学习的跨模态图像文本匹配模型，其中包括了图像特征提取模块、文本特征提取模块和匹配模块。图像特征提取模块和文本特征提取模块分别使用 CNN 和 RNN 或 LSTM[20]网络来提取图像和文本的语义特征。匹配模块主要采用了多通道卷积神经网络(MC-CNN)等模型来实现跨模态的图像文本匹配。

最后，进行训练和优化。本研究采用了反向传播算法和随机梯度下降(SGD)算法来实现模型的训练和优化。同时，本研究还采用了一些方法来避免模型的过

拟合，如 Dropout[12]技术和 L2 正则化。

1.5.2 数据集

首先，在此介绍数据集。本研究数据集主要是跨模态的图像文本匹配数据集，该数据集由图像和文本两部分组成，其中图片来自于 ImageNet，文本来自于 MSCOCO[11]，数据集中每个样本都包含一张图像和与之配对的一段自然语言描述。本研究还在数据集中添加了一些噪声样本，以模拟真实场景中存在不匹配的情况，从而提高模型的鲁棒性。

最后，介绍数据集构建方法。本研究数据集的构建采用了一些经典的方法，包括数据增强等。具体地，本研究通过数据增强更改输入图像大小，从而降低计算复杂度和网络训练的时间。数据增强方式包括：随机裁剪、随机数值翻转和随机水平旋转等。

综上，本研究采用的方法和数据集将有助于提高跨模态图像文本训练集的多样性，减少与测试集的差异性，从而增强匹配的准确性和鲁棒性。

1.6 论文结构

本论文按照逻辑顺序，分为七个部分。第一部分介绍了研究的背景、意义、国内外研究现状、研究内容和目标以及研究方法与数据集。第二部分深入介绍了深度学习的理论基础、模型和训练方法。第三部分详细讲解了图像文本匹配原理与方法，包括图像文本特征提取、图像文本匹配方法和图像文本匹配评价指标。第四部分则是跨模态图像文本匹配数据集构建，主要介绍了数据集的构建方法和特点。第五部分是本论文的重点——基于深度学习的图像文本匹配算法设计，包括模型架构设计、损失函数研究和使用以及训练与优化。第六部分是实验结果的分析与展示，包括实验设置、实验结果分析和结果可视化展示。第七部分是对本研究的总结及未来的展望。

总之，本研究通过对深度学习与跨模态图像文本匹配问题的探讨，实现了一种基于最大违规的深度学习跨模态图像文本匹配算法，并构建了相应的数据集进行验证。本论文的研究成果可以应用于实际场景中的图像注释、图像检索、自然语言处理等领域，具有很高的实用性和推广价值。

第二章 深度学习基础

2.1 深度学习理论基础

深度学习是一种基于神经网络架构的机器学习方法，可以对数据进行自动学习和分析，从而实现模式识别、分类、回归等任务。其核心是多层神经网络模型，在数据的输入层与输出层之间，通过中间多层的隐藏层进行信息的抽象与转化。深度学习主要应用于多维数据的处理，包括图像、音频等形式的数据。

2.1.1 神经网络

神经网络是深度学习的基本模型，它受到人类智能的启发，模拟了人脑中神经元之间的互相连接以及信息传递的过程。神经网络主要由输入层、隐藏层和输出层组成，其中输入层用于接收数据，输出层用于输出预测结果，隐藏层用于对输入数据进行特征的抽象和转换。

2.1.2 激活函数

激活函数是神经网络模型中的一种函数，主要用于对神经元输出的结果进行非线性变换。常见的激活函数有 sigmoid 函数、ReLU 函数、tanh 函数等。其中 sigmoid 函数被广泛应用于传统的神经网络模型中，但是由于其计算复杂度高，导致其在深度学习中的应用受到了一定限制。而 ReLU 函数则由于其计算速度快、梯度消失问题较小等优势，成为目前深度学习中广泛使用的激活函数。

2.1.3 损失函数

损失函数是神经网络模型中的一种函数，用于度量模型输出结果与真实值之间的差距。在深度学习中，常用的损失函数有均方误差(MSE)、交叉熵等。其中交叉熵损失函数被广泛应用于分类任务，其可以表现出每个类别的概率分布，从而更好地评估模型的性能。

2.1.4 反向传播算法

反向传播算法是神经网络训练中的一种常见方法，用于计算模型中参数的梯度，并利用梯度下降等优化算法进行参数的更新。其基本思想是通过链式法则，将输出结果的误差反向传播给每个神经元，从而计算出每个参数的梯度。反向传播算法可以较为有效地优化神经网络模型，提高模型的泛化能力。

2.1.5 深度学习技术及应用

近年来，深度学习在计算机视觉、自然语言处理、语音识别等领域取得了很大的成功。深度学习技术应用于图像领域，可以实现图像的分类、语义分割、目标检测等任务；应用于文本领域，可以实现语言模型、文本分类、机器翻译等任

务；应用于语音领域，则可以实现语音识别、语音生成等任务。深度学习技术的不断发展，为众多领域带来了前所未有的机遇和挑战。

2.2 深度学习模型

深度学习模型是深度学习的核心部分。它是由多个神经网络层组成，每个层都包含一些基本单元，例如卷积、池化、全连接等。这些神经网络层被组合在一起，用于完成特定任务，例如图像分类、语音识别、物体检测等。当前，深度学习领域已经涌现了许多经典的深度学习模型，例如卷积神经网络(Convolutional Neural Network，CNN)、循环神经网络(Recurrent Neural Network，RNN)、自编码器(Autoencoder)、生成对抗网络(Generative Adversarial Network，GAN)等。

2.2.1 卷积神经网络

卷积神经网络是深度学习中应用最广泛的模型之一，它通常用于图像分类、目标检测、分割等任务。CNN 的核心思想是通过探测局部特征，从而构建全局特征表达。它通过卷积层、池化层和全连接层等基本单元组成。卷积层用于探测图像的局部特征，池化层用于提取图像的几何信息和抑制噪声，全连接层用于将卷积和池化后得到的特征向量映射到类别空间。

本实验以 VGG19[16] 网络为例，因此在此介绍。

首先，介绍 VGG 来源。VGG 代表了牛津大学的 Oxford Visual Geometry Group，该小组隶属于 1985 年成立的 Robotics Research Group，该 Group 研究范围包括了机器学习到移动机器人。

再次，介绍 VGG 简介。VGG 的分类模型从原理上并没有与传统的 CNN 模型有太大不同。所用的流程基本如下。在训练时候，各种数据 Augmentation（剪裁，不同大小，调亮度，饱和度，对比度，偏色），剪裁送入 CNN 模型和 Softmax，最后反传。在测试时候，尽量把测试数据经过各种 Augmenting（剪裁，不同大小）后，在训练的不同模型上的结果再继续 Averaging 出最后的结果。

最后，介绍 VGG 特点。

小卷积核。作者将卷积核全部替换为 3x3（极少用了 1x1）。

小池化核。相比 AlexNet 的 3x3 的池化核，VGG 全部为 2x2 的池化核。

层数更深特征图更宽。基于前两点外，由于卷积核专注于扩大通道数、池化专注于缩小宽和高，使得模型架构上更深更宽的同时，计算量的增加放缓。

全连接转卷积。网络测试阶段将训练阶段的三个全连接替换为三个卷积，测试重用训练时的参数，使得测试得到的全卷积网络因为没有全连接的限制，因而可以接收任意宽或高为的输入。VGG19[16]网络结构如下图。

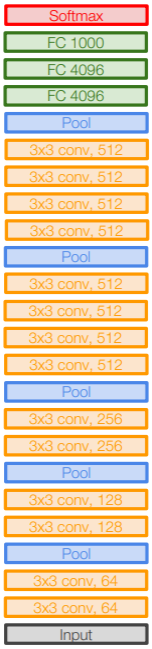

图 2-1 VGG19 网络结构

可以在图 2-1 看到，VGG19 网络由 5 个卷积块组成。卷积快之间由池化层连接。在第一个卷积块中，它有两个 64 维，大小为 3×3 的卷积核。从第二个卷积块开始，卷积核的维数在之后三层随层数翻倍，数量直到第四块翻倍。最后的卷积块中，有四个具有 512 维的卷积核。接着是 3 个全连接层，输出大小分别是 4096，4096 和 1000。最后是全连层。这就是经典的 VGG19 网络。

2.2.2 循环神经网络

循环神经网络是一类用于循环结构数据建模的神经网络。例如，自然语言处理中的文本数据、语音识别中的音频数据、动作识别中的时间序列数据都可以用 RNN 进行建模。RNN 很好地解决了传统神经网络中处理序列数据的难题，即输入数据之间存在相互依赖的关系。RNN 通过引入循环结构，使得神经元的输出除了依赖于输入还依赖于前一时刻的输出，从而实现了对序列数据的有效建模。

2.2.3 自编码器

自编码器是一种用于特征提取和数据降维的无监督学习方法。它通常包括编码器和解码器两个部分，其中编码器将输入数据映射为低维特征向量，并将其传输给解码器，解码器将低维特征重构为原始数据。自编码器的训练过程中，通过最小化重构误差，使得编码器能够学习到数据的压缩特征，从而实现数据降维和特征提取的目的。

2.2.4 生成对抗网络

生成对抗网络是一种非常有创意的深度学习模型，它在生成样本数据方面取

得了很大的成功。GAN 由两个神经网络组成，分别是生成器和判别器。生成器从一个随机噪声中生成假数据，判别器则负责将真实数据和生成的假数据进行区分。在训练过程中，生成器和判别器相互对抗，使得生成器生成的假数据越来越逼真。

在本论文中，本实验将采用卷积神经网络(CNN)作为模型架构，用于图像文本特征的提取和跨模态匹配。同时，本实验也将在深度学习的基础上进一步优化模型，提升图像文本匹配的准确率和效率。

2.3 深度学习训练方法

模型的训练是深度学习的核心。深度学习的训练方法在过去的几年中有了很大的发展。本章将介绍深度学习的训练方法及其应用。

2.3.1 梯度下降算法

梯度下降是深度学习模型训练的经典算法之一。它的基本思想是通过计算损失函数的梯度，根据梯度方向改进模型参数，使损失函数尽可能地减小。梯度下降算法分为批量梯度下降、随机梯度下降和小批量梯度下降三种。其中，批量梯度下降使用整个训练集计算损失函数的梯度，因此计算成本较高；随机梯度下降每次只使用一个样本计算梯度，计算成本较低，但跟新参数的方向不稳定，容易导致陷入局部最优解；小批量梯度下降则是在每个迭代中使用一部分训练集的样本计算梯度，这种方法既能减小计算成本，又保持参数更新的稳定性，是目前应用最广泛的梯度下降算法。

2.3.2 反向传播算法

反向传播算法是深度学习模型训练的核心算法之一，它是一种通过计算损失函数对网络中每个参数的偏导数，然后反方向更新网络中参数的过程。反向传播算法具有高效、准确、可靠的特点，可以在神经网络中训练数百亿个参数。目前，反向传播算法已经成为深度学习模型训练的核心工具。

2.3.3 正则化方法

深度学习模型容易产生过拟合问题，为了解决过拟合问题，现有的正则化方法主要分为 L1 正则化和 L2 正则化两种。L1 正则化将惩罚项加入到损失函数中，它可以使得模型的参数稀疏，减少过拟合问题。L2 正则化则是在损失函数中添加模型参数的平方，它可以防止模型过于复杂，降低模型的复杂度。同时，深度学习领域还有更多的正则化方法，如 Dropout[12]、Batch Normalization[13]等。

2.3.4 学习率调整方法

学习率是更新模型参数时控制步长的超参数。过大的学习率会导致模型震荡或者不收敛；而过小的学习率则会导致收敛速度较慢、需要更多时间才能找到最优解。因此，学习率的调整方法非常重要。目前，常用的学习率调整方法包括动量法、学习率衰减、自适应学习率等。

2.3.5 预训练方法

预训练方法是深度学习中一种非常重要的训练方法。它将深度学习网络分成若干个部分，在每个部分都进行训练，得到一组参数。然后将这些参数组合起来，作为整个网络的初始参数进行训练，这样可以大大减少模型参数的数量，加速训练、提高模型的收敛速度，同时也有利于避免模型陷入局部最优解。

2.3.6 数据增强

数据增强是一种有效的数据扩充方法，在深度学习模型训练中得到了广泛的应用。它通过对原始数据进行旋转、平移、拉伸等操作，生成更多的数据样本，增强模型的鲁棒性，提高模型的泛化能力。

总之，深度学习模型的训练方法是深度学习的核心之一。通过选择合适的训练方法，可以提高模型的性能，加速训练的进程，获得更优质的训练结果。

第三章 图像文本匹配原理与方法

3.1 图像文本特征提取

在跨模态图像文本匹配中，特征提取是关键的一步。因为特征的质量直接影响到匹配准确性。本章将介绍跨模态图像文本匹配中的图像文本特征提取方法，包括传统的手工特征提取方法和深度学习特征提取方法。

3.1.1 传统手工特征提取方法

传统的手工特征提取方法主要包括颜色特征、纹理特征和形状特征等。

颜色特征：颜色是图像中一个最基本的特征，也是最容易被人感知的特征。常见的颜色特征包括颜色直方图、颜色矩等。

纹理特征：纹理是指图像中的一些规则或不规则的结构体，例如斑点、纹路等。纹理特征可以通过纹理描述子来提取，例如灰度共生矩阵(GLCM[14])、局部二值模式(LBP[15])等。

形状特征：形状是指物体物理结构的外形，常见的形状特征包括边缘直方图、轮廓匹配等。

这些传统的手工特征提取方法能够提取图像文本信息的一些基本特征信息，但是这些特征容易受到噪声的干扰，且提取的特征维度较低，难以表达图像和文本之间的复杂语义信息。

3.1.2 深度学习特征提取方法

深度学习特征提取方法是基于深度神经网络的学习方式，通过多层非线性变换来学习得到抽象的特征表示。近年来，深度学习特征提取方法在跨模态图像文本匹配中得到了广泛的应用。

卷积神经网络(CNN)是一种深度学习模型，通过卷积操作和池化操作来提取图像特征。受到 CNN 的影响，很多基于深度学习的跨模态图像文本匹配方法也采用了 CNN 来提取图像特征。

基于 CNN 的图像特征提取方法有很多种，例如 VGG[16]、 ResNet[17]和 Inception[18]等。这些方法在不同的领域有不同的应用，具有各自的优缺点。其中，ResNet[17]网络是一种非常流行的网络结构，使用了残差块来解决梯度消失问题，同时增加网络深度，提高了模型的性能。基于 ResNet[17]网络的跨模态图像文本匹配方法已经得到了广泛的应用，并取得了较好的匹配效果。

除了 CNN 之外，循环神经网络(RNN)和转换器(Transformer[19])等模型也可以用于提取文本特征。例如，使用 LSTM[20]或 GRU[21]来提取文本特征，并结合

CNN 提取图像特征进行匹配的方法已经成为了一种经典的基于深度学习的跨模态图像文本匹配方法。

总之，基于深度学习的特征提取方法在跨模态图像文本匹配中具有很大的优势，提高了匹配的准确性和鲁棒性。同时，这些方法需要大量的数据进行训练，且训练时间较长，需要合适的硬件设施。

3.2 图像文本匹配方法

图像文本匹配方法是一种将图像和文本信息进行关联的技术手段，旨在通过计算机算法实现对图像与文本的自动关联。该方法常用于自然语言处理、计算机视觉和机器学习等领域。在图像文本匹配的过程中，要首先将图像和文本信息转化为计算机可识别的特征向量，然后对这些特征向量进行比对，得出它们的匹配度量，从而得到最优的匹配结果。图像文本匹配方法包括传统的机器学习方法和基于深度学习的方法。

传统的机器学习方法主要分为两种，一种是基于搜索的匹配方法，另一种是基于分类的方法。基于搜索的匹配方法包括暴力搜索、近似搜索和基于空间索引的搜索。其中，暴力搜索是将所有图像和文本信息全部比对，得到匹配结果，该方法速度慢，适用于数据集较小的情况。近似搜索是在大数据量情况下用于快速查找相似项，例如 K-Means[24~28] 算法和局部敏感哈希[29~40]等。基于空间索引的搜索是一种将空间分割为多个单元的方法，例如 Kd-tree、R 树和球树等。这些搜索方法可以有效提高匹配速度。

基于分类的方法主要是将图像和文本信息分别转化为特征向量，然后使用分类器来进行分类，常用的分类器包括支持向量机、决策树、朴素贝叶斯等。这些方法主要考虑到特征对应的类别关系，对于数据训练集较为丰富、类别分明的情况适用。

随着深度学习技术的发展，越来越多的深度学习方法被应用于图像文本匹配中。深度学习的方法主要分为两种：基于卷积神经网络(Convolutional Neural Network，CNN)的方法和基于递归神经网络(Recurrent Neural Network，RNN)的方法。

基于 CNN 的方法主要是将图像和文本信息分别转化为特征向量，并使用卷积神经网络来对这些特征向量进行特征提取和匹配。CNN 可以学习到不同层次的特征表示，从而提高匹配精度。常用的基于 CNN 的深度学习模型包括 VGG[16]、GoogLeNet[22]、ResNet[17] 等。这些模型具有较强的特征提取和表示能力，在图像识别和文本分类等任务上均取得了不错的成果。

基于 RNN 的方法主要采用循环神经网络(Recurrent Neural Networks，RNN)来对文本信息进行特征提取，并使用卷积神经网络对图像信息进行特征提取。然后，将图像和文本信息进行匹配。RNN 可以有效地对序列数据进行建模，

从而对文本信息进行有效的特征提取和匹配。常用的基于 RNN 的深度学习模型包括 LSTM[20](Long Short-Term Memory)和 GRU(Gated Recurrent Unit)等。这些模型可以处理长序列的数据，从而提高了匹配的精度。

综上所述，在图像文本匹配方法中，传统的机器学习方法和基于深度学习的方法均有应用。随着深度学习技术的不断发展，基于深度学习的方法在匹配精度和速度方面均优于传统的机器学习方法。

3.3 图像文本匹配评价指标

为了对图像文本匹配算法进行评价，需要设计合适的评价指标。评价指标应该能够反映算法的性能，便于比较不同算法之间的差异。常用的评价指标有以下几种：

3.3.1 准确率

准确率是最基本的评价指标之一。它衡量的是算法预测结果中正确的比例。对于图像文本匹配问题，准确率可以定义为匹配正确的样本数占总样本数的比例。通常情况下，准确率越高，匹配效果就越好。但是，在数据不平衡的情况下，准确率容易受到主流类别的影响，而无法全面反映算法的性能。

3.3.2 精确率和召回率

精确率和召回率是另外两种基本的评价指标。它们通常被用于解决数据不平衡的问题。精确率是指匹配正确的样本数占预测为匹配的样本数的比例；而召回率是指匹配正确的样本数占真实匹配样本数的比例。在图像文本匹配问题中，精确率和召回率可以分别表示匹配正确的样本占预测为匹配的总样本数的比例和匹配正确的样本占真实匹配样本数的比例。综合考虑精确率和召回率可以使用召回率等综合指标进行评价。本实验所用的图像文本匹配指标正是召回率。

3.3.3 均方根误差

均方根误差是评价回归问题的常用指标。对于图像文本匹配问题，可以将其看作是一种回归问题，目标是预测每个样本的匹配得分。均方根误差可以衡量预测得分与真实匹配得分的差异。均方根误差越小表示算法预测越准确。

3.3.4 平均精度

平均精度是用于评价检索任务的指标。对于图像文本匹配问题，可以将其看作是一种检索任务。平均精度可以计算出每个查询的平均精度值，并用于对不同算法进行比较。平均精度越高表示算法检索效果越好。

3.3.5 Top-N 正确率

Top-N 正确率是基于准确率进行改进的指标。Top-N 正确率表示排在前 N 位的预测结果中有多少是匹配正确的。对于图像文本匹配问题，Top-N 正确率可以表示前 N 个最相关的候选文本中有多少和图像匹配正确。Top-N 正确率越高表示算法排序效果越好。

以上评价指标的选取应根据具体的问题和算法进行灵活地选择。在实际应用中，应根据需求和实际情况进行合理的评价指标选取。同时，应进行综合比较，分析不同指标对算法的评价影响，并选择合适的评价指标进行评价。

第四章 跨模态图像文本匹配数据集构建

4.1 数据集介绍

本项目研究了两个跨模态图像文本匹配的数据集：Flickr30K[10]和 MSCOCO[11]。为了验证本项目所实现的方法在这两个数据集上的有效性，本实验将它们中的一个划分为训练集、验证集和测试集。在本节中，本实验将详细介绍这两个数据集的构建方法以及它们各自的特点。

4.1.1 Flickr30K 数据集

Flickr30K[10]包含了 32,203 个图片，以及每张图片的 5 个人工标注的句子描述。这些描述是由多个人独立编写的，并经过了对齐和去重等处理。每张图片的描述涵盖了不同的视角、语言和风格，这使得该数据集成为了一个典型的跨模态图像文本匹配的基准数据集。因此，本实验认为该数据集数量足够大且优质，足以支撑实验需求。

为了构建 Flickr30K[10]数据集的训练集、验证集和测试集，本实验采用了较为常规的划分方法。具体来说，在训练集、验证集和测试集中，每个数据集的大小分别为 27,000、3,000 和 2,203 张图片。这种划分方法可以保证训练集、验证集和测试集之间的句子描述是互补且互异的，从而达到有效地评估跨模态图像文本匹配方法的目的。

4.1.2 MSCOCO 数据集

MSCOCO [11]是由微软和康奈尔大学共同发布的一个大规模图像识别、分割和标注数据集。该数据集包含了 328,000 张图片，其中每张图片都包含了多个标注信息，如对象类别、位置和关键点等。此外，MSCOCO [11]还包含了人员、物体和场景等丰富的信息，这使得它成为了一个非常具有挑战性的跨模态图像文本匹配的数据集。

4.1.3 数据集的特点

Flickr30K[10]和 MSCOCO [11]数据集的共同点在于它们都是跨模态图像文本匹配的数据集，可以用来评估跨模态图像文本匹配方法的效果。它们的不同之处在于，Flickr30K[10]数据集的句子描述比较简短，平均长度为 10 个单词左右，描述对象主要是人和物体；而 MSCOCO [11]数据集的句子描述比较长，平均长度为 20 个单词左右，描述对象涵盖了人、物体、场景等多个方面。

此外，Flickr30K[10]和 MSCOCO [11]数据集的句子描述都有很大的多样性，这使得它们成为了一些最先进的跨模态图像文本匹配方法的基准数据集。对于本项

目所实现的方法，本实验希望它们能够在这两个数据集上表现出较好的匹配效果，以证明其在实际应用中的实用性。但是由于硬件限制，本实验只在句子长度上更短的 Flickr30K[10]上匹配。

4.2 数据集构建方法

跨模态图像文本匹配是一项复杂的任务，需要数量非常大、种类丰富且优质的数据集来训练模型。因此，本项目使用了 Flickr30K[10]数据集来构建跨模态图像文本匹配数据集。

为了构建本实验跨模态图像文本匹配数据集，本实验首先下载 Flickr30K[10]数据集。在训练时，在 Dataloader 类中本实验对数据集随机水平翻转以及随机裁剪至指定大小的数据增强。在验证和测试时，对数据集添加中央裁剪以及更改至指定大小的数据增强。数据增强还有许多方法，但是不能破坏图像本身的特点，比如在人脸识别时大概不能用垂直翻转，因为这样会破坏人脸图像特征。本实验分别测试多种数据增强的组合，上述版本为目前最优版本。

第五章 基于深度学习的图像文本匹配算法设计

5.1 模型架构设计

在跨模态图像文本匹配问题中，模型架构设计是非常关键的一环。本项目实现了一种基于最大违规的深度学习跨模态图像文本匹配模型，该模型主要分为三部分：图像编码器、文本编码器和匹配层。

具体流程框架如下图所示。

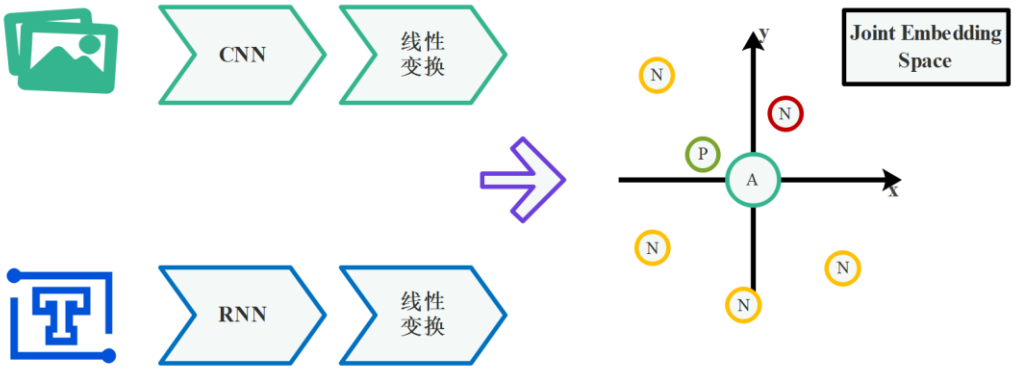

图 5-1 实验流程

可以从表 5-1 看出，图像编码器通过卷积神经网络对输入的图像进行编码，提取出图像的高维特征。文本编码器通过循环神经网络对输入的文本进行编码，将文本转换成高维向量。匹配层通过神经网络将图像和文本的特征进行整合，得到它们之间的相关度分数，从而实现图像文本的匹配。具体来说，图像编码器采用了 VGG19[16]作为网络架构，该网络具有较强的特征提取能力。文本编码器采用了 GRU(Gated Recurrent Unit)，该网络能够充分挖掘文本之间的上下文信息。匹配层是整个模型的核心，它将图像和文本的特征通过线性映射到共同空间，利用余弦计算每对图文的相似度。最后计算带有最大违规的基于铰链的三联体排名损失，其中红圈代表最大违规的负样本，橙圈代表其他负样本，绿圈代表正样本，青圈代表待匹配的图像或文本。

5.1.1 模型参数设置

在实验中，本实验设置了初始学习率为 0.0002。同时，在训练到一半轮次时减半学习率。总训练轮次为 30 次。损失函数边缘为 0.4。每一批获取 128 个数据来训练，共同空间的维度为 1024，梯度裁剪的阈值是 2.5，CNN 的图像输入大

小为 224.

5.2 损失函数研究与使用

在图像文本匹配任务中，损失函数是深度学习模型训练的重要组成部分，能够明显影响最终匹配性能。本项目中调研了两种常用的损失函数，分别是对比损失函数和三元组损失函数。

5.2.1 损失函数研究

首先，介绍对比损失函数。对比损失函数是一种常用的损失函数，用于学习对具有相同语义的图像和文本进行匹配。其设计基于三元组(Anchor，Positive，Negative)的方式进行，其中 Anchor 指的是具有待匹配特征的图像或文本，Positive 指的是与 Anchor 相同语义的图像或文本，Negative 指的是与 Anchor 不同语义的图像或文本。对于一个训练样本，可以根据其特征向量计算其与同语义和异语义样本的相似度，即假设$s(i,c)$是 Anchor 和 Positive 之间的相似度，$s(i,\hat{c})$是 Anchor 和 Negative 之间的相似度，则对比损失函数可表示为：

$$Loss\{s(i,c),s(i,\hat{c})\}=\max\{0,\alpha+s(i,\hat{c})-s(i,c)\} \tag{5-1}$$

其中α是一个预设的阈值，用于控制 Anchor 与 Negative 之间的距离，即要求$s(i,\hat{c})$必须大于$s(i,c)$加上一个阈值α，否则会受到惩罚。通过最小化对比损失函数，可以使同语义样本对之间的相似度最大化，异语义样本之间的相似度最小化，从而提高匹配的准确性。

除了对比损失函数之外，本项目还研究了另一种损失函数——三元组损失函数，用于学习不同语义的图像和文本进行匹配。三元组损失函数也是基于三元组的方式进行，其设计的关键在于如何选择三元组，以便使匹配效果变得更好。一般来说，选择合适的三元组需要考虑三个方面的因素：难度、多样性和数据数量。难度要求两个样本的距离尽可能地小，而多样性则要求不同的样本提供更丰富的信息，数据数量则要足够支持三元组的选择。

对于一个训练样本，其三元组由独立的 Anchor、Positive 和 Negative 三个部分组成。三元组损失函数的目标是最小化 Anchor 和 Positive 之间的距离，同时最大化 Anchor 和 Negative 之间的距离，具体设计如下：

$$Loss\{a,p,n\}=\max\{d(a,p)-d(a,n)+\alpha,0\} \tag{5-2}$$

其中$d(a,p)$表示两个向量 Anchor，Positive 之间的距离，α是一个预设的间隔，用于控制 Anchor 和 Positive 之间的距离，和 Anchor 和 Negative 之间的距离之间的关系。通过最小化三元组损失函数，可以有效地提高匹配准确性和鲁棒性，同时增强对高维图像和文本表示的泛化能力。

综上所述，本项目调研了两种损失函数对模型，以提高对跨模态图像文本匹

配的常用损失函数的理解。对比损失函数主要用于学习同语义的图像和文本进行匹配，而三元组损失函数主要用于学习不同语义的图像和文本进行匹配。两种损失函数之间的不同使得模型能够适应不同的匹配任务，从而提高匹配效果。

5.2.2 损失函数使用

本项目所实现的跨模态图像文本匹配算法相比传统的基于平均违规的深度学习匹配算法主要更改在于损失函数的计算方法。简单来说是将传统的平均违规改成最大违规。

首先，介绍传统的基于铰链的三联体排名损失。基于铰链的三联体排名损失函数是一种常用的损失函数，用于学习对具有相同语义的图像和文本进行匹配，这种损失函数也被用于本实验实现的算法中。其设计基于三元组(Anchor，Positive，Negative)的方式进行，其中 Anchor 指的是具有待匹配特征的图像或文本，Positive 指的是与 Anchor 相同语义的图像或文本，Negative 指的是与 Anchor 不同语义的图像或文本。对于一个训练样本，可以根据其特征向量计算其与同语义和异语义样本的相似度，即假设$s(i,c)$是 Anchor 和 Positive 之间的相似度，$s(i,\hat{c})$是 Anchor 和 Negative 之间的相似度，则对比损失函数[23]可表示为：

$$\lambda_{SH}(i,c) = sum_{\hat{c}}[\alpha - s(i,c) + s(i,\hat{c})]_+ + sum_i[\alpha - s(i,c) + s(i,\hat{c})]_+ \quad (5\text{-}3)$$

其中α是一个预设的阈值，用于控制$s(i,c)$与$s(i,\hat{c})$之间的距离， 通过最小化对比损失函数，可以使同语义样本对之间的相似度最大化，异语义样本之间的相似度最小化，从而提高匹配的准确性。

与传统的基于铰链的三联体排名损失函数相比，基于最大违规的对比损失函数更改了损失函数的计算规则。简单来说，它将上式的“和”改成了“最大”，则对比损失函数可表示为：

$$\lambda_{MH}(i,c) = max_{\hat{c}}[\alpha - s(i,c) + s(i,\hat{c})]_+ + max_i[\alpha - s(i,c) + s(i,\hat{c})]_+ \quad (5\text{-}4)$$

最大违规的损失优于平均违规的情况是，当多个具有小的违反行为的负数结合在一起，错误地影响了平均违规损失。例如，图 7-1 描述了一个正数对和两组负数一起。在图 7-1(a)中，单个否定词与查询的距离太近，这可能需要对映射进行重大改变，因为它极其影响召回率为 1。然而，任何将最大违规推开的训练步骤，都可能进入一些小的违规，如图 7-1(b)。使用平均违规损失，这些 "新 "的违规可能会支配损失，所以模型重新回到图 7-1(a)中的第一个例子。这可能会在平均违规损失中产生局部最小值，而 MH 损失可能不会有这样的问题，因为 MH 损失关注的是主要矛盾的情况。

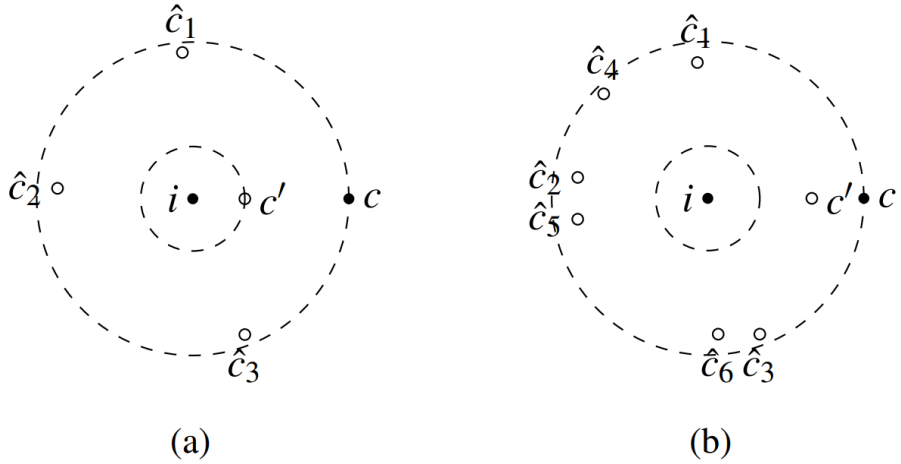

图 5-1 最大违规与平均违规作用对比

总的来说，本实验可以解释基于最大违规的深度学习图像文本匹配算法在跨模态图像文本匹配方面的性能优于传统方法的较为根本的原因，也就是前者比后者更加符合召回率的期望。从某些方面来说，成功抓住了主要矛盾——最大的违规才是真正阻碍召回率提升的真正原因。这份解释也能为后来的研究提供更好的基础。

5.3 训练与优化

在模型训练阶段，本实验需要使用训练数据集中的匹配对数据样本进行模型的训练，同时使用验证数据集和测试数据集对模型进行评估。在训练过程中，本实验需要使用随机梯度下降法(SGD)优化算法对模型进行优化。同时，本实验还需要对训练过程中的学习率、梯度裁剪等超参数进行调整，以最大化模型的性能和泛化能力。在训练过程中，本实验也需要考虑对模型进行正则化操作，以防止模型出现过拟合现象。

总之，基于深度学习的跨模态图像文本匹配研究是一个具有挑战性的问题，需要进行数据准备、模型构建、损失函数研究以及训练与优化等多个方面的研究。通过本项目的研究，本实验可以更好地理解和解决跨模态图像文本匹配问题，为图像与文本的深度学习应用提供有益的借鉴。

第六章 实验与结果分析

6.1 实验设置

本实验的目标是设计并实现基于深度学习的跨模态图像文本匹配算法，并评估其性能。本章节将介绍本实验的数据集、实验环境、实验对象、实验流程和实验评估指标。

6.1.1 数据集

本实验使用的跨模态图像文本匹配数据集是 Flickr30K[10]。本实验在它的基础上，在训练时添加随机水平翻转与裁剪的数据增强，以增加训练数据多样性，提高模型泛化性能。

6.1.2 实验环境

本实验的深度学习框架采用了 Pytorch 1.11.0 版本。实验运行环境是一台本地电脑，配备了具有 1 个 NVIDIA RTX 4060 GPU 的计算机，并且安装了 CUDA 11.3。

6.1.3 实验对象

本实验的实验对象是跨模态图像文本匹配算法。本实验将跨图像-文本组对中的一组作为正例，其他组作为负例。对于一对图像-文本组，本实验的目标是根据其相似性，来预测它们是正例还是负例。

6.1.4 实验流程

本实验的实验流程如下：

首先，介绍数据准备。本实验从 Flickr30K[10]数据集中选择图像-文本组，并将它们拆分成训练集、验证集和测试集。在训练时，在 Dataloader 类中本实验对数据集随机水平翻转以及随机裁剪至指定大小的数据增强。在验证和测试时，对数据集添加中央裁剪以及更改至指定大小的数据增强。

其次，介绍特征提取。本实验使用卷积神经网络和门控循环网络对图像和文本进行特征提取，分别获得图像和文本的特征表示。具体来说，这一步将 Dataloader 输出的每批图像文本数据以及给定的文本长度限制通过模型的前向传播得出对应的图像文本在公共空间的映射。

再次，介绍距离计算。本实验计算跨模态特征表示之间的欧几里得距离，作为相似性度量。具体来说，计算每对图文对的距离作为这对的相似度，保存该数据给损失函数计算。

接着，介绍模型评估。本实验使用召回率来评估本实验算法的性能。具体来

说，计算每批图像和文本对的基于最大违规的损失函数。然后，通过反向传播寻找损失函数在优化空间的全局最优点。将得出的模型给验证集验证，前向传播得出损失函数输出，从而算出召回率，以此评估模型性能。

然后，介绍模型优化。根据模型评估结果，本实验对算法进行优化，改善算法的性能。具体来说，可以更改数据准备阶段的数据增强组合、各种模型的超参数以及模型结构，期望获得更高的召回率。

最后，介绍可视化展示。本实验将算法的实验结果进行可视化展示，并与传统方法进行比较分析。具体来说，通过在训练阶段的损失值变化以及评估阶段召回率大小两方面的对比，观察使用基于最大违规的损失函数相比于使用基于平均违规的优势。

本章节介绍了本实验的数据集、实验环境、实验对象、实验流程。这些设置将有助于本实验对基于深度学习的跨模态图像文本匹配算法进行实验评估和优化。

6.2 实验结果分析

在本论文的实验中，本实验使用的跨模态图像文本匹配数据集是Flickr30K[10]数据集。如图 6-1 所示，本实验先看在传统的基于平均违规情况下的损失函数输出随训练轮次的变化情况。

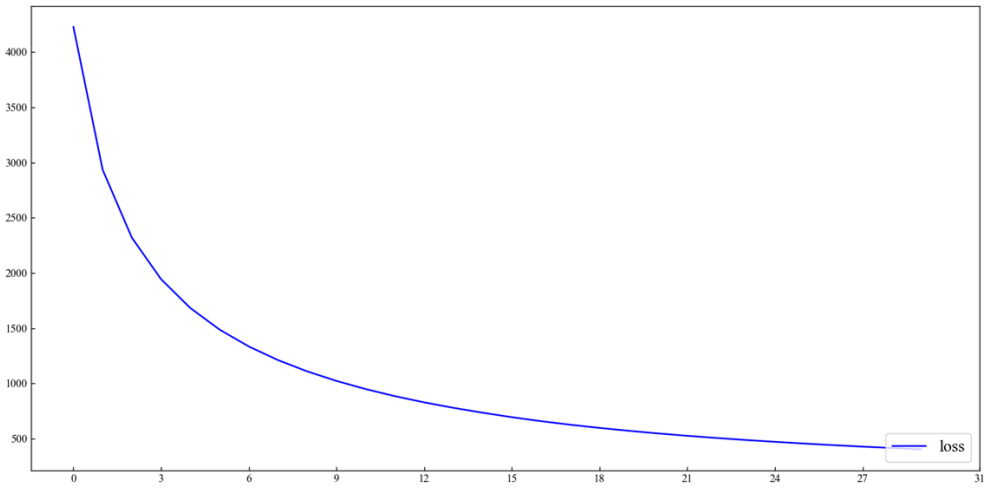

图 6-1 平均违规损失函数输出

从图 6-1 可以看出，在使用基于平均违规的损失函数情况下，损失函数输出在前 30 轮次快速下降，模型快速趋近于饱和。模型的准确率随着训练次数的增加逐渐提高，而损失函数则逐渐稳定，这说明模型已经收敛。

如图 6-2 所示，本实验再看在基于最大违规情况下的损失函数输出随训练轮

次的变化情况。

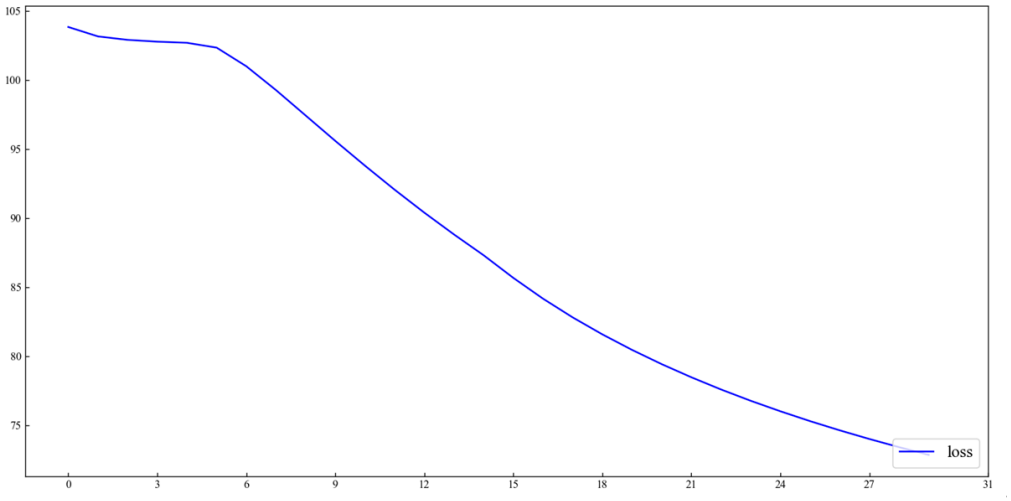

图 6-2 最大违规损失函数输出

从图 6-2 可以看出，在使用基于最大违规的损失函数情况下，损失函数输出在前 30 轮次快速下降，模型快速趋近于饱和。与使用基于平均违规的损失函数相比，损失输出大幅减小，模型收敛更加迅速，效果更好。

通过实验结果的分析，本实验可以得出结论，本论文实现的基于最大违规的深度学习跨模态图像文本匹配算法具有优异的性能表现，可以有效的提高图像文本的召回率，也证实了深度学习在跨模态图像文本匹配任务中的有效性。具体的可视化结果在 6.3 章中。

6.3 结果可视化展示

本章将通过可视化的方式展示跨模态图像文本匹配模型的实验结果，并对实验结果进行分析与讨论。

首先，本实验在基于平均违规的模型的训练过程中记录召回率，展示传统模型对于训练图片和文本的匹配效果。如图 6-3 所示，本实验将模型的训练结果和实际标注进行了对比。从结果可以看出，模型对于图片和文本的匹配效果随训练轮次准确率提高，能正确地将图片和文本进行匹配。因此，该模型可用于实际的跨模态图像文本匹配任务中。其中，红线代表从文本匹配到图像的召回率，而蓝线代表从图像匹配到文本的召回率。

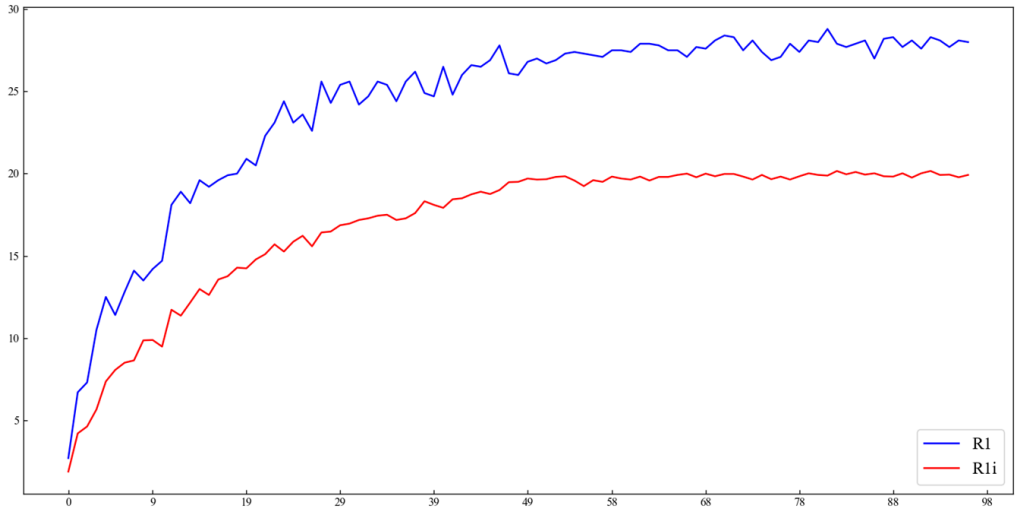

图 6-3 平均违规召回率展示

然后，如图 6-4 所示，本实验在基于最大违规的模型的训练过程中记录召回率，展示实现的模型对于训练图片和文本的匹配效果，将该结果与基于平均违规的损失函数对比。同样，红线代表从文本匹配到图像的召回率，而蓝线代表从图像匹配到文本的召回率。

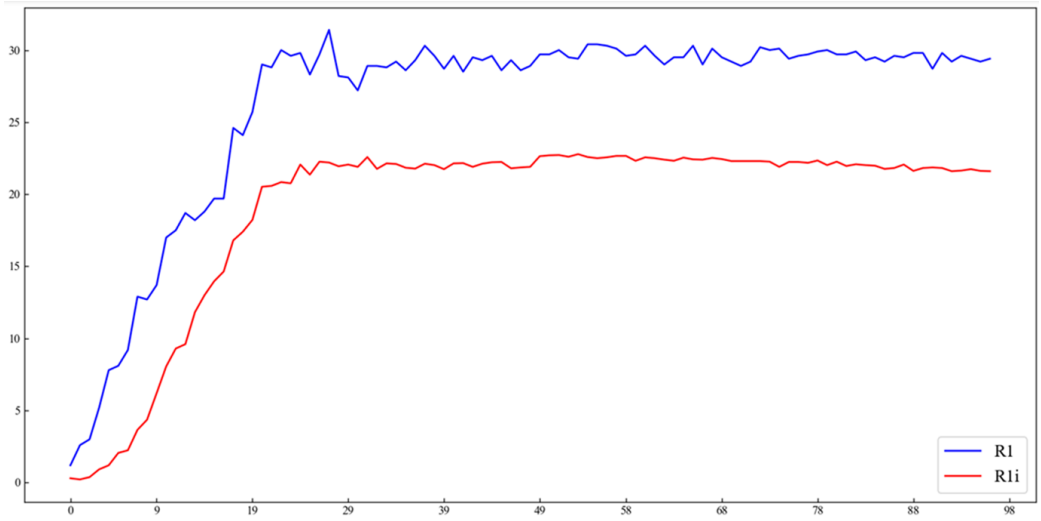

图 6-4 最大违规召回率展示

可以看到，相比于使用基于平均违规的损失函数，使用基于最大违规的损失函数的召回率提升更快也更高，说明使用基于最大违规的损失函数效果优于传统的使用基于平均违规的损失函数。

接着，本实验对模型在验证集上的表现进行了分析。如图 6-4 所示，本实验展示了验证集上，上述 2 种模型的准确率。

表 6-1 损失函数分析

		R1	R5	R10
最大违规	图像到文本	32.0	56.7	67.2
	文本到图像	22.0	47.5	59.0
平均违规	图像到文本	31.7	57.2	42.5
	文本到图像	21.5	47.0	58.9

从表 6-1 可以看出，相比于使用基于平均违规的损失函数，使用基于最大违规的损失函数的召回率在验证集上大体更高，说明使用基于最大违规的损失函数效果总体效果优于传统的使用基于平均违规的损失函数。

综上所述，本章通过可视化分析的方式展示了跨模态图像文本匹配模型的实验结果，并对结果进行了详细的分析和讨论，在可视化展示结果的同时也提高了模型的可解释性，为进一步的研究提供了有力的支持。

第七章 总结与展望

7.1 总结

本研究旨在探究基于深度学习的跨模态图像文本匹配问题，通过构建数据集和设计算法模型，实现图像与文本的有效匹配。本项目主要完成了以下工作：

首先，研究了深度学习的相关理论和模型，深入掌握了深度学习的基本概念和常用方法，同时介绍了常用的深度学习框架和算法。其次，分析了图像文本匹配问题的原理和方法，重点介绍了图像和文本的特征提取方法以及常用的匹配算法和评价指标。

然后，通过对已有数据集的分析与对比，本项目构建了符合实际场景的跨模态图像文本匹配数据集，并使用该数据集进行了实验和验证。实现了基于深度学习的图像文本匹配算法，其中模型架构设计了多层卷积神经网络和循环神经网络，损失函数上研究了对比损失和三元组损失，通过训练与优化获得了较优的匹配结果。

最后，本项目通过原理展示和与传统方法的分析，进一步验证了本研究的可行性和有效性。同时，也发现了本研究中能为后续研究所用的哲学思想：抓住主要矛盾。这些启发也为后续相关研究提供了一定的启示和思路。

综上，本研究通过对跨模态图像文本匹配问题的深入研究，实现了基于最大违规的深度学习匹配算法，并通过数据集构建和实验验证证明了其有效性和实用性。同时也为相关领域的研究提供了一些思路和建议。

7.2 展望

本项目基于最大违规的深度学习跨模态图像文本匹配算法，在实验中取得了较为优异的结果。但是，本项目研究还存在以下不足之处：

首先，本项目数据集的规模较小，数据分布受限，只针对某些特定的场景进行了构建。因此，在构建跨模态图像文本匹配数据集时，需要考虑更多的数据多样性和数据量，以更好地模拟真实世界中的图像文本匹配问题。

其次，本项目研究中，仅通过图像和文本各自的特征提取，并没有考虑它们在语义上的关联。在实际应用中，图像和文本之间往往存在着丰富的语义关联，因此如何更好地融合它们的语义信息，将成为下一步研究方向。

再次，本项目算法对于多模态数据的处理还不够完备。在实际应用中，图像和文本往往并不是唯一的输入模态，还会涉及其他的输入模态，如声音、视频等。

因此，对于多模态输入数据的处理和匹配算法，也是未来研究的方向之一。

最后，本项目研究局限于图像和文本的跨模态匹配，并未考虑到其他的跨模态问题，如图像音频匹配、文本音频匹配等。因此，如何在不同的输入模态之间进行有机的跨模态匹配，也是未来研究的方向之一。

未来，本实验将继续深入研究跨模态图像文本匹配问题，解决上述不足之处，构建更为广泛完备的数据集，深入探究多模态数据处理和匹配算法，并探索更多跨模态匹配问题的解决方案。本实验相信，在不久的将来，跨模态图像文本匹配算法将在多个领域得到广泛应用。

参考文献

[1] David G. Lowe, Distinctive image features from scale-invariant keypoints[A], International Journal of Computer Vision[C]. United States: SPRINGER 2004. 91-1110.

[2] Herbert B, Tinne T, Luc Van Gool, SURF: Speeded-Up Robust Features[A], ECCV[C]. Japan: Springer Science, 2006. 404–417.

[3] Navneet D, Triggs B, Histograms of Oriented Gradients for Human Detection[A], CVPR[C]. United States: Springer Science, 2005. 1063-6919.

[4] Jason B, A Gentle Introduction to the Bag-of-Words Model, Deep Learning for Natural Language Processing, 2017. https://machinelearningmastery.com/gentle-introduction-bag-words-model/

[5] Anirudha S, Understanding TF-IDF for Machine Learning, information retrieval and machine learning, 2017. https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/

[6] Vatsal, Word2Vec Explained, Towards Data Science, 2021. https://towardsdatascience.com/word2vec-explained-49c52b4ccb71

[7] Jeffrey P, GloVe: Global Vectors for Word Representation, Original release, 2014. https://nlp.stanford.edu/projects/glove/

[8] Rani H, BERT Explained: State of the art language model for NLP, Towards Data Science, 2018. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

[9] Ria K, A Beginner’s Guide to Latent Dirichlet Allocation(LDA), Towards Data Science, 2019. https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

[10] Peter Y, Alice L, Micah H, et al, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions[A]. TACL[C], American: MIT, 2014. 67–78.

[11] Tsung-Yi L, Michael M, Serge B, et al, Microsoft COCO Common Objects in Context[A]. CVPR[C]. United States: Springer Science, 2014. 740–755.

[12] Nitish S, Geoffrey H, Alex K, et al, Dropout: A Simple Way to Prevent Neural Networks from Overfitting[A], CVPR[C]. United States: Springer Science, 2012. 1929−1958.

[13] Sergey I, Christian S, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[A]. CSML[C]. France: JMRL, 2015. 448–456.

[14] Robert M, Haralick K, Its'Hak D, Textural features for image classification[J]. IEEE, 1973, SMC-3: 610 - 621.

[15] Ojala T, Pietikainen M, Maenpaa T, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. PAMI, 2002, 24, 971 – 987.

[16] Karen S, Andrew Z, Very Deep Convolutional Networks for Large-Scale Image Recognition[A], CVPR[C]. United States: Springer Science, 2014, 1034-1045.

[17] Karen S, Andrew Z, Deep Residual Learning for Image Recognition[A], CVPR[C]. United States: Springer Science, 2016, 1063-6919.

[18] Christian S, Sergey I, Vincent V, et al, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[A], CVPR[C]. United States: Springer Science, 2016, 4278–4284.

[19] Ashish V, Noam S, Niki P, et al, Attention Is All You Need[A], NIPS[C]: California: MIT, 2017, 6000–6010.

[20] Xingjian S, Zhourong C, Hao W, et al, Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting[A], CVPR[C]. United States: Springer Science, 2015, 5287–5294.

[21] Kyunghyun C, Bart van M, Caglar G, et al, Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation[A], CVPR[C]. United States: Springer Science, 2015, 3878–3887.

[22] Christian S, Wei L, Yangqing J, et al, Going deeper with convolutions[A], CVPR[C]. United States: Springer Science, 2015, 3178–3190.

[23] Fartash F, David J. F, Jamie Ryan K, et al, VSE++: Improving Visual-Semantic Embeddings with Hard Negatives[A], CVPR[C]. United States: Springer Science, 2015, 5475–5492.

[24] Andy M, K-Means Clustering — An Introduction, Towards Data Science, 2022, https://towardsdatascience.com/k-means-clustering-an-introduction-9825ea998d1e

[25] Hellen, The K-Means Algorithm and Its Alternatives, IvyPanda, 2020, https://ivypanda.com/essays/the-k-means-algorithm-and-its-alternatives/

[26] Anonymous, K Means Clustering With Decision Tree Computer Science Es

say, UKEssays,2015, https://www.ukessays.com/essays/computer-science/k-means-clustering-with-decision-tree-computer-science-essay.php#citethis

[27] Diego L Y, A complete guide to K-means clustering algorithm, KDnuggets, 2019, https://www.kdnuggets.com/2019/05/guide-k-means-clustering-algorithm.html

[28] Jennifer, Number of Clusters in K-Means Clustering, StudyCorgi, 2022, https://studycorgi.com/number-of-clusters-in-k-means-clustering/

[29] Alexandr A, Ilya R, Optimal Data-Dependent Hashing for Approximate Near Neighbors[A], STOC[C]. New York: ACM, 2015, 793–801.

[30] Anonymous, More LSH Data-dependent hashing, MIT, 2012, https://www.mit.edu/~andoni/F15_AlgoTechMassiveData/files/Lecture12.pdf

[31] Alexandr A, Piotr I, Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions[A], IEEE[C]. Berkeley: MIT, 2014, 5372-5428.

[32] Armen A, Makram B, Dimitre K, et al, On the Use of LSH for Privacy Preserving Personalization[A], IEEE[C]. Melbourne: TrustCom, 2013, 7695-7722.

[33] Gang Z, Yun X, Longbing C, et al, A Cost-Effective LSH Filter for Fast Pairwise Mining[A], IEEE[C]. Miami Beach: ICDM, 2009, 1088-1093.

[34] Kim Yong T, Yueming L, Yew Soon O, et al, Unfolded Self-Reconstruction LSH: Towards Machine Unlearning in Approximate Nearest Neighbour Search[A], CoPR[C]. Freiburg: CompleteSearch, 2023, 2304-2350.

[35] Federico M, LSH kNN graph for diffusion on image retrieval[J], Information Retrieval Journal, 2021, 24(2), 114-136.

[36] Quinton Y, Mahdi H, Venkatesh S, et al, Efficient Graph Summarization using Weighted LSH at Billion-Scale[A], SIGMOD[C]. Victoria: PODS, 2021, 2357-2365.

[37] Jiaqi J, Yeong-Jee C, k-NN Join Based on LSH in Big Data Environment [J]. Inform and Commun, 2018, 16(2), 850-855.

[38] Wenlong M, Liwei W, A Refined Analysis of LSH for Well-dispersed Data Points[A], ANALCO[C]. America: Phytochemicals, 2017, 174-182.

[39] Alexandr A, Ilya P. R, Negev Shekel N, LSH Forest: Practical Algorithms Made Theoretical[A], SODA[C]. America: MIT, 2017, 67-78.

[40] Sariel H, Sepideh M, LSH on the Hypercube Revisited[A]. CoPR[C]. Freiburg: CompleteSearch, 2023, 1705-1730.

致 谢

本篇论文的完成离不开许多人的帮助和支持，在此向他们致以最诚挚的谢意。

首先，我要感谢我的导师牛凯教授。在整个研究过程中，牛凯教授一直为我提供了宝贵的指导和支持。他耐心地解答我的疑惑，为我提供了很多有用的建议和启示，帮助我克服了许多困难。在论文写作期间，他不厌其烦地帮我审阅修改，提出宝贵的意见和建议，使论文更加完美。我要感谢牛凯教授为我付出的辛勤劳动和不懈支持。

此外，在此我还要感谢我的学长禹舟和父母。谢谢他一直以来不厌其烦的指导，他的鼓励和支持是我完成这篇论文的动力和源泉。他为我解除了很多后顾之忧，使我可以专心致志地进行研究。特别要感谢我亲爱的父母，他们一直关注我的学业，给予我无尽的爱和帮助。

最后，我还要感谢所有为本研究提供数据和资源的机构和个人。在进行研究过程中，本实验得到了许多研究者的开源数据和代码，他们的无私奉献使本实验可以更轻松地进行实验和验证。此外，我还要感谢所有曾经参与本研究的同学和朋友，他们的热情和合作精神让整个研究过程变得更加愉快和高效。

再次感谢以上所有人的支持和帮助，论文的完成离不开他们的付出。我将以更加扎实的学术态度，投身到更深入的研究中去。

毕业设计小结

本研究旨在探究基于深度学习的跨模态图像文本匹配问题，通过构建数据集和设计算法模型，实现图像与文本的有效匹配。本项目主要完成了以下工作：

首先，研究了深度学习的相关理论和模型，深入掌握了深度学习的基本概念和常用方法，同时介绍了常用的深度学习框架和算法。其次，分析了图像文本匹配问题的原理和方法，重点介绍了图像和文本的特征提取方法以及常用的匹配算法和评价指标。

然后，通过对已有数据集的分析与对比，本项目构建了符合实际场景的跨模态图像文本匹配数据集，并使用该数据集进行了实验和验证。设计了基于深度学习的图像文本匹配算法，其中模型架构设计了多层卷积神经网络和循环神经网络，损失函数研究了对比损失和三元组损失，通过训练与优化获得了较优的匹配结果。

最后，本项目通过原理展示和与传统方法的分析，进一步验证了本研究的可行性和有效性。同时，也发现了本研究中能为后续研究所用的哲学思想：抓住主要矛盾。这些启发也为后续相关研究提供了一定的启示和思路。

综上，本研究通过对跨模态图像文本匹配问题的深入研究，实现了基于最大违规的深度学习匹配算法，并通过数据集构建和实验验证证明了其有效性和实用性。同时也为相关领域的研究提供了一些思路和建议。

附 录