基于深度多模态检索系统设计

摘要

随着互联网的迅速发展，不同模态的媒体数据也在迅速增长。传统检索通过为数据赋予标签来实现检索功能，其存在新模态数据标签难以获取、查询模态受到限制等一系列问题。在此背景下，用于检索多模态数据的基于深度学习的跨模态检索技术迅速受到关注。而在跨模态检索领域中，最常见的任务是对文本和图像之间的跨模态检索。为此，本文对基于深度学习的文本-图像生成二值化查询向量跨模态检索方法进行了研究，并提出了一种基于深度学习的跨模态检索系统，本文的主要工作如下：

1．对跨模态检索相关方法进行了研究，分别考察了跨模态检索模型的构建方法、跨模态检索方法中图像特征提取方法、文本特征提取方法和跨模态三元组损失策略，并对相关技术进行了介绍。

2．本文提出了一种基于深度学习的跨模态检索系统设计方法，其使用双编码器跨模态检索模型结构，使用
ResNet18 和 Bert12 分别为图像和文本进行特征提取，并使用 Dense 进行共同表示学习。在训练中，本文提出使用经过

修改的跨模态三元组损失函数进行训练，并提出了跨模态检索系统的使用流程。

3．实验中使用 MSCOCO2017 对本文所提方法进行训练，在使用文本作为查询向量时，其检索速率达到 9.457query
/s，HitRate@N 和 Precision@N 分别达到了 69.58%、84.24%、88.68%、69.58%、68.18%和 67.92%，其中 N 分别取 1，5，
10 。实验证实了该方法的有效性。关键词：跨模态检索、特征提取、深度学习、哈希

1 绪论

1.1 题目背景及目的

随着互联网的迅速发展，不同类型的媒体数据也在迅速增长，例如文本、图像、视频和音频等等，在这种背景下，用于检索多模态数据的跨模态检索技术正越来越受到关注。通过跨模态检索，用户可以通过提交某一种模态的数据，来查询另一种模态的对应数据。例如，用户可以通过一段音频信息，来查询数据库中符合音频描述的图片信息。从结果上看，跨模态检索主要可以分为两个类别：生成实值检索向量的跨模态检索方法与生成二值化检索向量的跨模态检索方法。前一种方法能够获得更高的精度，但是需要消耗更大的储存空间和花费更长的检索时间，而在跨模态检索对查准率的需求是低于查全率和检索速度的。在跨模态检索领域中，最常用的是对文本与图像之间跨模态检索的研究。因此在本文中，我们将探讨生成二值化检索向量的图像与文本之间的跨模态检索方法，并据此设计一个文本检索图片的深度多模态检索系统。

1.2 国内外研究状况

1.2.1 跨模态检索研究现状

在此之前，已经有许多学者从不同的方向提出了关于提取不同模态之间相似性的跨模态哈希方法[1]。在这些跨模态哈希算法中，一般首先使用特征抽取方式获得该模态数据的抽象化表达，然后将这些被抽取出来的特征向量映射到不同模态的公共汉明空间中，在之后的检索中，衡量不同模态在汉明空间中的距离，以确定其相似度，从而达到跨模态检索的效果。在较早的时候，学者们一般使用较浅的特征提取方法，如基于相关语义最大化(SCM)[2]、集体矩阵分解哈希(CMFH)[3]、CMSSH[4]等方法，这些方法都取得了一定的效果。近年来，随着深度学习模型的特征提取能力越来越强，越来越多学者使用深度学习模型来代替较早的特征提取网络，如[5, 7, 8 , 11, 12]，等方法都取得了很好的结果。Wang. 等人[5]使用 AlexNet[6]作为图片特征提取网络，Cheng. 等人[7]使用 LSTM 作为文本的特征提取网络，Lu. 等人[8]使用了 BERT[9]和 Faster-RCNN[10]作为特征提取网络等等，这些特征提取能力越来越强的网络让跨模态检索方法的成功率上升了一个台阶。Zhan. 等人[11]提出了关注不同模态内的语义信息之间的匹配，后续学者们进一步提出了使用能够抽取语义信息的特征提取[8]、使用注意力机制关注两模态之间的相关语义信息等等方法、Wei. 等人[12]提出了深度语义匹配来提升跨模态检索的效果等等。大部分的 Image2Text 跨模态检索模型使用两个编码器分别对图片和文本进行编码，Lu. 等人[13]提出了使用一个基于 transformer[14]的模型来学习不同模态的信息，这种方法使用了大量预训练模型的知识，比双编码器的模型取得了更好的结果。Wang. 等人[15]提出了使用 Graph Neural Network(GNN)来建立不同对象间的关系。

1.2.2 图像特征提取网络研究现状

在跨模态检索中具有图片作为某一模态信息时，一般会使用图像特征提取网络来提取该模态信息。比较常见的
有[6, 16, 17, 18]等等，文献[6, 16]使用不同层数的 Convolutional Neural Network(CNN)提取图像特征。随着
CNN 堆叠层数的加深学者们发现出现了梯度消失等现象，于是提出了使用残差层防止梯度消失[17]。Newell. 等人
[18] 提出 Stacked Hourglass Model(SHM)，其通过堆叠多个 Hourglass Model 来强化网络对于多尺度特征的提取能力。

在跨模态检索领域中，学者们一开始单纯使用特征提取网络来提取整体特征。Wang. 等人[5]使用 AlexNet 提取特征，Xi. 等人[19]使用 ResNet 提取特征。随后提出了使用具有局部语义提取能力的模型进行特征提取。

Karpathy. 等人[20]使用 Region Convolutional Neural Network(RCNN)[21]来为图像进行编码，随后 He. 等人[22] 提出了使用 Faster-RCNN[10]来提取图像特征，并利用 Region of Interest(ROI)提取图片元素的语义信息，此外

Xi. 等人[19]为图像卷积部分增加由注意力机制产生的掩膜，以期望提取模态间共享的语义信息。

1.2.3 文本特征提取研究现状

文本特征提取是图片-文本双编码器跨模态检索中的一个重要问题。常见的本文特征提取方法首先将文本转为单词语词向量[23, 24]等等，文本转化为词向量后通过特征提取方法提取出文本特征向量。Cheng. 等人[7]提出了使用 Long Short-Term Memory(LSTM)来处理，该模型能在提取过程获得时序信息。随着深度学习的火爆，具有时序记忆功能的 Recurrent Neural Network(RNN)[25]被提出，并且结合了神经网络的 LSTM 以及其变体 GRU[26]等被提出。 LSTM、GRU 等长短程记忆网络与 RNN 相比，具有更强的长程时序记忆功能，Schuster. 等人[27]提出了构建双向时序神经网络，启发了后续 BiRNN、BiGRU 等双向时序神经网络的提出，文献[28, 15]等方法提出使用 Graph Neural Network(GNN)进行自然语言处理，Luong. 等人[29]提出了为模型添加注意力机制，这些方法都在文本特征提取中取得了成功。

随着计算机能力的发展，Skip-Gram[30]、GloVe[31]等方法提出了 Pretrained Language Models(PLM)结构，该

结构通过大型语料库进行预训练，学习通用语言表示形式，以帮助进行下游的自然语言处理任务(NLP),但是这种方法无法捕获上下文中的高级概念。随后 ELMo[32]、BERT[9]等具有强大表达能力的 PLM 模型被提出，BERT 提出使用不同数量的 Transformer[14]堆叠组成模型的编码器和解码器，该方法提出后在自然语言处理的各个领域取得了巨大的成功，随后 ERNIE[33]、GPT-3[34]等强力的 PLM 模型也被提出。

1.3 研究内容和论文构成

本文通过阅读和总结深度跨模态检索文献，对深度跨模态检索方法开展了初步研究和进行了实验仿真，并设计了一个基于 PLM 和 ResNet 的深度跨模态检索系统。该跨模态系统是一个双编码器跨模态检索系统，使用预训练模型

ResNet18[17]作为图像编码器、预训练模型 BERT[9]作为文本编码器、经过修改的 Triplet Ranking Loss[35]作为损失函数、MSCOCO 数据集[36]作为进行实验的数据集。

随后在第二章我们介绍深度跨模态检索的相关技术，第三章设计一个基于深度学习的多模态检索系统，第四章对本文设计的深度多模态检索系统进行实验和结果分析，随后我们将对本次研究内容进行总结。

2 深度多模态检索相关技术介绍

深度多模态检索的主要任务是使用某一种类型的数据作为查询去检索其他类型的相关数据[A Comprehensive

Survey on Cross-modal Retrieval]。在具体实现上，使用特征提取网络提取不同类型数据的特征，然后通过共同表征学习(Common Representation Learning)学习不同类型数据的共同表示方法，接着通过与共同表征学习结合的编码器为数据库中的数据生成跨模态索引，然后通过查询数据生成的查询向量与数据库中生成的索引计算排名，得到与查询数据相似度最高的前 N 个数据库数据作为检索的返回。多模态检索系统的结构如图 2.1。在本文中，我们仅关注文本和图片两种类型数据的跨模态检索，使用 Bert 和 ResNet 分别作为文本和图片的特征提取网络，使用全连接

神经网络生成二值化编码。

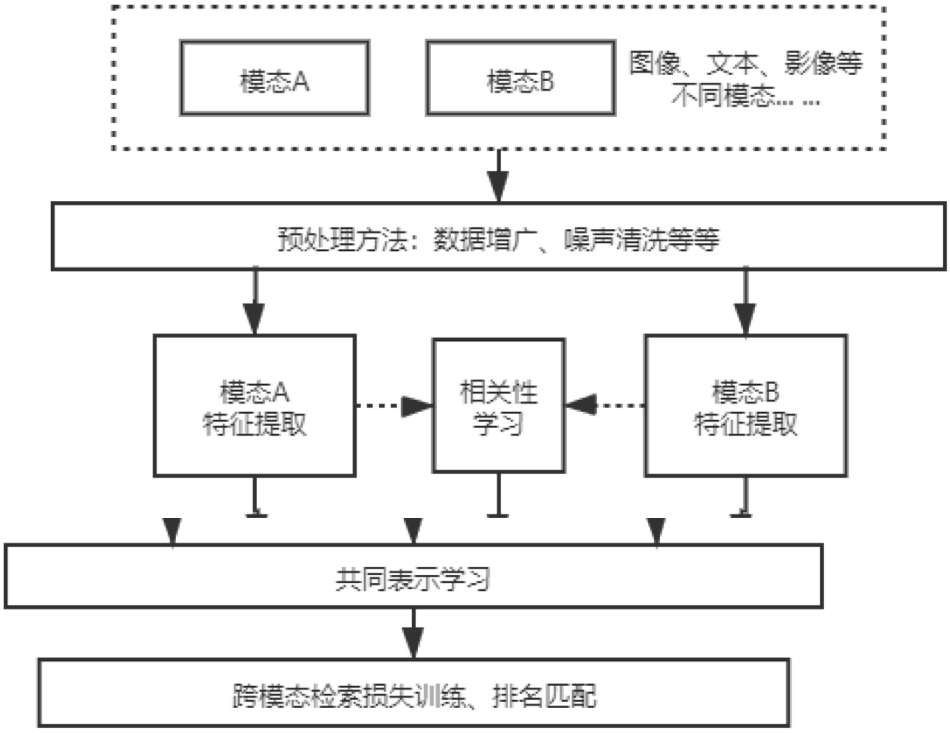


图 2.1 基于深度学习的跨模态检索模型的简单介绍图

在接下来的三节中，本文将分别介绍多模态检索系统的文本特征提取(2.1 节)、图片特征提取(2.2 节)相关技术，以及介绍在跨模态检索中常用的三元组损失函数(2.3 节)。

2.1 跨模态检索中的自然语言处理相关技术

在本次跨模态检索任务中，文本作为跨模态检索的查询数据，需要具有较强的实时性。经典循环神经网络，如 RNN、LSTM 等模型，在前向传播与反向传播时需要按顺序输入词向量并进行，无法充分理用 GPU 的并行计算加速作用［25］，如图 2.2(a)，并且由于结构受限，只能沿着时间单元反向传播，多次堆叠会导致反向传播距离过深而梯度消失，并且难以使用残差结构缓解梯度消失现象，如图 2.2(b)。BERT 由 Transformer 多层次堆叠而成，每个层次中通过自注意力机制和位置嵌入向量，使得词向量可以在每个层次中并行输入，并且每个 Transformer 的全连接层前后具有残差结构，可以缓解深层堆叠神经网络训练时梯度消失现象，此外，Transformer 的多头注意力机制与自注意力机制的组合能够提供强大的特征提取能力。因此，在本次任务中使用具有自注意力机制和残差机制的 Transformer 输入输出并行堆叠的 BERT 模型作为文本的特征提取网络，然后使用单层全连接神经网络和二值化网络进行共同表征学习并最终生成文本二值化查询码。

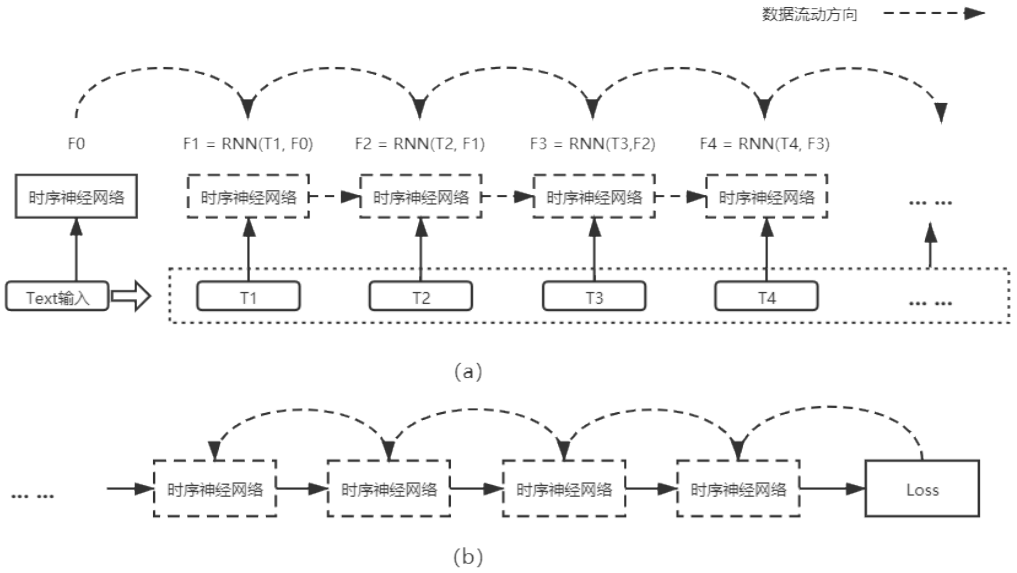


图 2.2 时序神经网络计算顺序

2.1.1 多头自注意力机制

在前后向量具有依赖性的时序数据中，每个时间节点对数据表达的信息的贡献与其他相互关联的时间节点有关，因此可以通过所有时间节点的信息为每一个时间节点赋予不同的权值，其最早由机器翻译领域研究者提出为注

意力机制。其在深度学习模型使用中见公式(2.1)、(2.2):

$$\alpha = Activation(Wf + bias) \quad (2.1)$$

$$f_{out}^{i} = a_i \times f^{i} \quad (2.2)$$

其中 Activation 是深度学习激活函数，bias 与 W 是前馈神经网络中的偏置和权重矩阵，

$f = \left\{ f^{i} \right\}, i = 1,...n$ 为输入的词向量序列，n 为词向量序列长度。

多头自注意力机制属于注意力机制的一种，最早由文献[14]提出，为了充分考虑句子之中不同词语之间的语义及语法联系，自注意力机制考虑了每一个词与所有其他词之间的关联性并设置权重，使用加权结果作为该词语义的表达，如图 2.3。

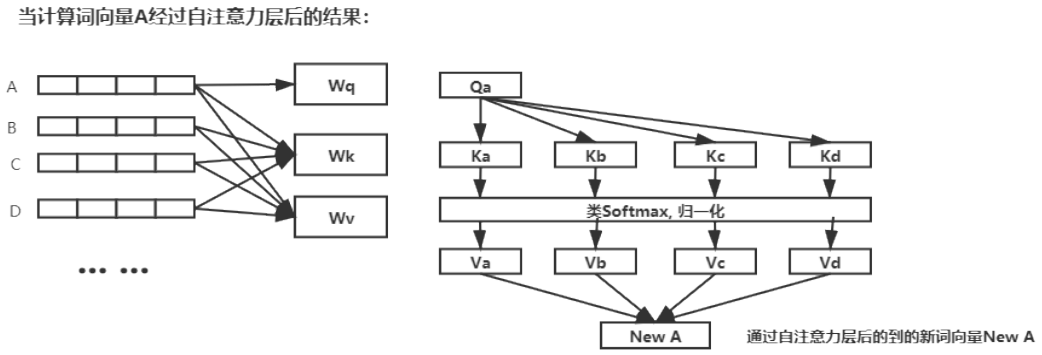


图 2.3 自注意力层词向量计算方法图

在计算时，首先创建W_k，W_q和W_v向量，每个向量大小都是$\langle N_x, N_k \rangle$,其中词向量为X_i，长度为N_x，则有：

$$Q_i = X_i W_q \quad (2.3)$$

$$K_i = X_i W_k \quad (2.4)$$

$$V_i = X_i W_v \quad (2.5)$$

$$Score_i = \sqrt[N_k]{\{Q_1 K_1, \ldots Q_n K_n\}} \quad (2.6)$$

$$\alpha_i = Soft\max\left(Score_i\right) \quad (2.7)$$

则第 i 个单词经过自注意力机制的表达为：

$$X_i^{new} = \sum_{i=1,\ldots n} \alpha_i V_i \quad (2.8)$$

此时每个新生成的词向量的长度为原长度的 ，则经过自注意力层后词向量序列变为 $\left\{X_1^{new}, X_2^{new}, \ldots\ldots, X_n^{new}\right\}$，其中 n 为词向量序列长度。 N_k/N_x

多头自注意力机制在计算时创建N_k/N_x个W_k，W_q，W_v向量，并每个都进行以上自注意力机制计算，将每个头生成的新的词向量拼接起来，则每个词向量长度与输入多头自注意力机制前相同。设置多头自注意力机制可以允许模型在不同的表示子空间里学习到相关的信息[14]。

2.1.2 Transformer 结构

Transformer 由两个子结构构成，分别是多头自注意力结构和简单的全连接神经网络。词向量首先输入带有残差连接的多头自注意力结构，随后输出的每个词向量分别输入带残差连接的全连接层中，两个子结构后都跟随着一个

层次归一化(Layer Normalization)层,其表示见公式(2.9),(2.10)：

$$X_{self} = LN(MulitHeadAttention(X) + X) \quad (2.9)$$

$$X_{out} = LN(Dense(X_{self}) + X_{out}) \quad (2.10)$$

其中 LN 为 Layer Normalization 层，MulitHeadAttention 为多头自注意力层，Dense 为前馈神经网络。

在 Transformer 中每个词向量并行输入，为了保留语序信息，在词向量输入时为每个词向量进行位置编码，嵌入

向量计算方法为：

$$PositionCode(POS,i) = \begin{cases} \sin(POS/10000^{2i/d_{model}}) & \mod(i,2)=0 \\ \cos(POS/10000^{2i/d_{model}}) & \mod(i,2)=1 \end{cases} \quad (2.11)$$

其中 POS 为词向量中元素位置，d_{model}为词向量维度，i 代表词向量序列中的第 i 个词向量。

最后通过将词向量和位置编码求和得到输入 Transformer 的向量。

Transformer 结构如图 2.4。

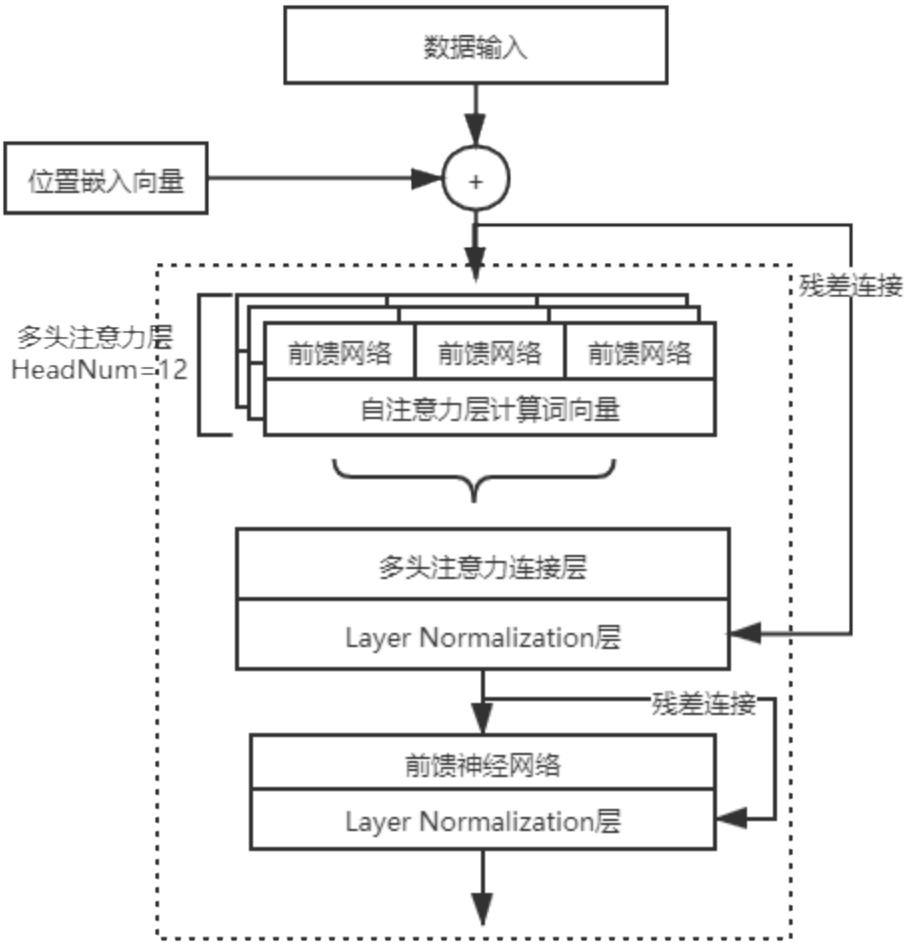


图 2.4 单层 Transformer 结构图

2.1.3 跨模态检索文本编码器

在跨模态检索任务中，通过文本编码器将文本编码为二值化查询向量，用于后续在共同表达索引中检索其他模态数据。文本首先使用通过词表进行分词和转为为词向量,然后对词向量序列进行位置编码，随后将进行位置编码后的向量输入到自然语言处理模型中，将自然语言处理模型的输出输入到全连接神经网络中进行维度变换，最后通过

非线性函数 Tanh 对模型的输出进行非线性变换。文本编码器可以表示如下：

$$X_{text\_vector}=Vocab(text) \tag{2.12}$$

$$X_{input}=PositionCode(X_{text\_vector})+X_{text\_vector} \tag{2.13}$$

$$X_{feature}=Bert12(X_{input}) \tag{2.14}$$

$$HashCode=Tanh(Dense(X_{feature})) \tag{2.15}$$

$$ByteCode_i=\begin{cases}-1 & HashCode_i<=0\\ 1 & HashCode_i>0\end{cases} \tag{2.16}$$

其中 Vocab 代表文字序列到词向量序列的词表映射。文本编码器的模型图如图 2.5 所示。

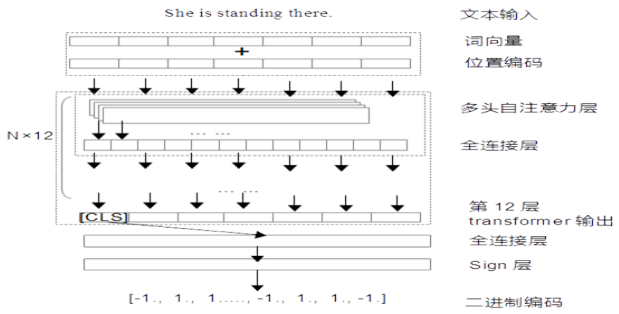


图 2.5 文本编码器模型图

2.2 跨模态检索中的图像特征提取相关技术

传统图像特征提取方法一般使用颜色直方图、纹理等方法，这些方法的优点是速度较快，但是需要手工设计特征，并且在图像类型复杂的情况下作用不够明显。相对而言，深度学习无需手工设置特征，并且更适合在复杂图像类型中进行特征提取。卷积神经网络使用卷积计算对输入信息进行平移不变分类，这保持了输入信息的空间结构信息，因此很适合用于进行图像特征提取。为了避免深度学习中网络加深出现的梯度消失现象，He. 等人[17]提出残差结构应用于深度神经网络。为了对海量不同环境下的图片数据进行特征提取，跨模态检索方法使用的图像特征提取网络应该具有强大的特征抽取和表征能力。深度学习中层数加深可以增强网络表征特征的能力，因此本次任务使用 ResNet 作为图像特征提取网络的骨架网络，并在骨架网络后连接全连接神经网络和 Tanh 非线性激活层。本文使用到的图像特征提取相关技术介绍如下。

2.2.1 卷积神经网络

卷积神经网络(Convolutional Neural Networks, CNNs)可以从大规模数据中分层学习其抽象特征，学习的特征具有较强的泛化性。对于基础 CNN 而言，一般可以分为卷积、激活和池化三个部分。通过将基础神经网络堆叠就可以实现特征提取的作用，在其输入连接其他模块以调整输入结构，并在输出连接其他网络以进行下游任务。

2-D 卷积是 CNN 中最基础的操作。具体到图像特征提取任务中，对于不同的颜色通道或特征通道，卷积核只能在同一通道中的 x,y 轴上滑动，而不能在通道轴上进行位移。因此进行基础 CNN 操作时，卷积核的深度必须要与通道的

深度一致，以确保能够对不同的通道特征进行提取。在实际进行卷积操作中，输入的特征尺寸为 $<C_{in},h,w>$、卷积核尺寸是 $<C_{out},C_{in},p,q>$,其权重为ω，输入的特征值为v，则卷积操作见公式(2.17)：

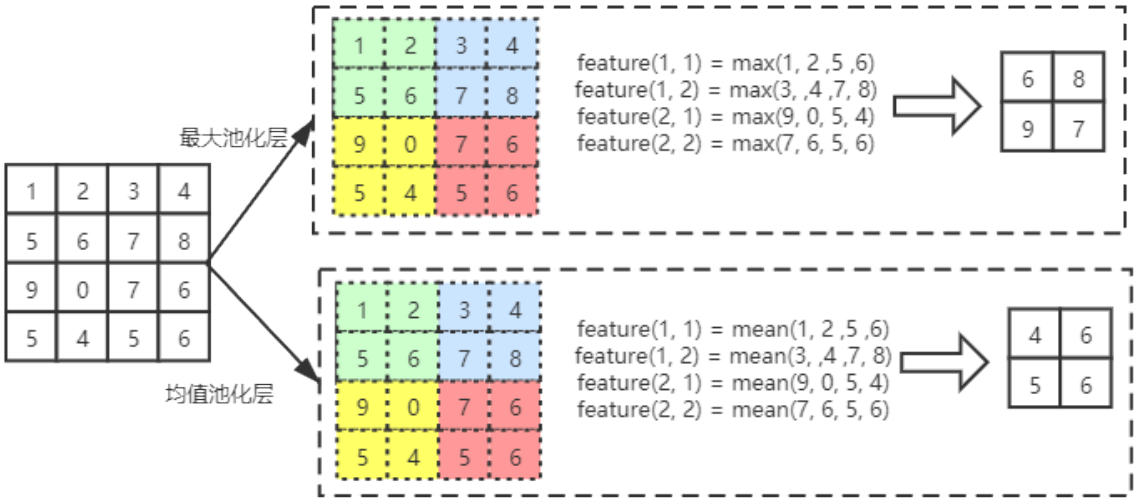
$$Conv_{x,y}=\sum_{i}^{p\times q}\omega_i v_i \tag{2.17}$$

输出的特征尺 $<C_{out},h,w>$,其中 C_{out} 寸为代表卷积核个数，每个卷积核与 o 的卷积结果作为输出特征的其中
一个通道，令输出 o_{out} 。 结构为 在进行卷积操作之后，通常会引入偏置和非线性激活函数，此操作用于使卷积操作获得非线性表达能力。常用的非线性激活函数有 ReLU、Sigmoid、Tanh。令激活函数为 H,其与偏置结合后表达如下：

$$v_{conv}=Activate(v_{out}+bias) \quad (2.18)$$

其中 bias 代表偏置值，其只与不同的卷积核相关。

为了改变特征图的尺寸，通常在网络中添加跨步卷积和池化层等方法。添加池化层即在卷积神经网络进行非线
性操作后进行池化炒作，对于一个池化尺寸是 的池化层，大小为 的特征图被输入到池化层中将被分割为 $<c,h/N,w/N>$ 个区域，并对每个区域进行池化操作。常见的池化操作有均值池化、最大值池化等等，池化后特征图的尺寸变为 $<c,h/N,w/N>$ 。池化的操作步骤如图 2.6。



$<N,N>$ $<c,h,w>$

图 2.6 池化层计算示例图

除了使用池化之外，卷积神经网络也通常使用跨步卷积层改变特征图尺寸和去除沉余。跨步卷积即对特征图进行步长为 N(N>1)的卷积操作,卷积结束后特征图长宽变为原来的 1/N。跨步卷积相较于池化层可以保留更多的特征，但是也会出现跟多参数和花费更多的计算量。

2.2.2 残差神经网络

VGG[16]、googlenet 试图通过加深深度学习网络深度以提高网络整体特征提取和表征能力，从而发现了基础深度卷积神经网络在到达一定深度后堆叠层数不能带来性能提升，反而使网络的收敛变得更慢，下游任务的效果也变得更差。受启发于计算机视觉领域的 VLAD[37],He. 等人[17]提出了深度残差网络并用于图像分类证明了其解决层数堆叠导致性能下降的问题的能力。深度残差网络提出了引入残差映射代替基础 CNN 中几个堆叠层直接拟合的基础映射。
令基础映射为 $H(x)$,残差映 $F(x)$ 射为：

$$F(x)=H(x)-x \quad (2.19)$$

转为为残差映射后，原 $H(x)$ 始映射可以表示为：

$$H(x)=F(x)+x \quad (2.20)$$

其中 x 代表原始特征输入。使用残差映射代替基础映射后获得残差模块，其表现为在基础 CNN 模块中增加短路连接，表示如图 2.7。