班　级　1513018

学　号　15130188018

西安电子科技大学

本科毕业设计论文

题　　目　基于多任务学习的文本表示方法研究

学　　院　计算机科学与技术学院

专　　业　软件工程

学生姓名　孙天祥

导师姓名　邱锡鹏 副教授

摘要

近年来，随着数据量的增加、算力的增强以及算法的成熟，以神经网络为代表的深度学习为自然语言处理领域带来了深刻的变革。在自然语言处理任务中，文本表示对模型性能起着关键作用。

然而，现实场景中常常因为文本标注难度大、成本高等原因导致训练数据有限，进而导致一般的深度学习算法难以学到泛化的表示。本文试图探索多任务学习在文本表示上的应用，以期与近期取得巨大成功的迁移学习范式互为补充，进一步提升神经网络模型的泛化能力。

目前，多任务学习方法已经被广泛地用在了卷积神经网络和循环神经网络两种主流的结构中，在计算机视觉和自然语言处理领域的多个任务上取得了成功。在本文中，我们在一种新的网络结构 – Transformer – 上探索了多任务学习的应用。首先，我们验证了两种传统的多任务学习架构在 Transformer 的应用，然后，我们针对 Transformer 的特点提出了两种逐层分化的多任务学习架构。在多个文本分类任务上的实验表明，四种多任务模型的准确率较单任务学习模型均取得了提升，其中，我们的逐层分化架构在四种架构中取得了最高的分类准确率。

最后，我们对模型进行了实例分析，总结了现有工作，并展望了未来多任务学习在自然语言处理以及深度学习的发展。

关键词：　自然语言处理　表示学习　多任务学习

ABSTRACT

In recent years, with the increase of data volume, the enhancement of computing power and the maturity of learning algorithms, deep neural networks has brought profound changes to the field of natural language processing. In natural language processing tasks, text representation plays a key role in model performance.

However, in real scenes, due to the difficulty and high cost of text annotation, training data is often limited, which makes it difficult for deep learning algorithms to learn a well-generalized representation. This paper attempts to explore the application of multi-task learning in text representation, in order to complement the paradigm of transfer learning, which achieved a great success recently, to further improve the generalization ability of deep neural networks.

At present, multi-task learning methods have been widely used in the two common networks, convolutional neural networks and recurrent neural networks, in the field of computer vision and natural language processing. However, this paper explores the application of multi-task learning on a new network, Transformer. First, we verify the application of two traditional multi-task learning architectures in Transformer. Then, considering the characteristics of Transformer, we propose two new architectures which are *layerwise differentiated*. Experiments on multiple text classification tasks show that the accuracies of our four multi-task models are consistently higher than single-task learning model. Particularly, our proposed *layerwise differentiated* architecture achieves the highest classification accuracy among the four architectures.

Finally, we conduct case study for our proposed model, summarize the existing works, and look forward to the future development of multi-task learning in natural language processing and deep learning.

Key words: **Natural Language Processing** **Deep Learning** **Multi-Task Learning**

目录

第一章 引言

本章首先介绍基于深度学习的自然语言处理的研究背景，阐述在该背景下多任务学习的研究价值及意义。接着简要介绍自然语言处理和多任务学习的研究进展及现状，并指出目前存在的问题与空缺。然后介绍本文的研究内容及目标，并概括了本文工作的创新之处。本章的最后给出了论文的主要内容和章节安排。

1.1 研究背景及意义

1950 年，阿兰·图灵（Alan Turing）提出了著名的图灵测试[1]，直接推动了人工智能从哲学探讨的层面上升到科学研究。随后不久，在 1956 年举办的达特茅斯会议上，人工智能（artificial intelligence，AI）的概念被正式提出，John McCarthy 将这一新兴领域的研究目标定义为：“让机器的行为看起来像人类所表现出来的智能行为一样”。自 1956 年至今的六十余年中，研究人员尝试了多种方法来实现这一愿景，人工智能领域也随着这些方法的成功与失败经历了数次热潮与低谷。近年来，随着数据量的增加和算力的增强，以神经网络为代表的深度学习异军突起，在语音识别[1][2]、计算机视觉[3][4][5] 等众多应用场景中取得了巨大突破，再次掀起了人工智能的研究热潮。

自然语言处理（natural language processing, NLP）被很多人认为是“人工智能皇冠上的明珠”，致力于使计算机具备理解和生成人类语言的能力。随着深度学习技术的发展，越来越多的研究者开始使用深度学习的技术解决自然语言处理中的难题[6][7][8][9]，在很多任务上远远超越了之前的传统方法。

然而，像绝大多数领域一样，对 NLP 领域的研究常常被划分为多个任务，如命名实体识别、阅读理解、机器翻译等。目前一般的做法是为当前关注的某一个任务设计特定的神经网络模型，在此单一任务及数据集上进行训练。遗憾的是，这样设计出来的模型常常具有较大的局限性，在某个数据集上表现优秀的模型可

[1]图灵测试是指，一个人在不接触对方的情况下，通过某种方式和对方进行一系列的问答。若在相当长时间内，他无法根据问答的情况判断对方是人还是机器，那么可以认为该机器具备智能。

能在另一个数据集上就会表现很差，即使这两个数据集来自同一任务。并且，由于 NLP 任务及数据集众多，一时之间各种神经网络结构层出不穷。深度学习刚刚帮助人们从“特征工程”中脱离出来，很快又陷入了所谓的“网络结构工程”。同时，这也将模型限制在了特定领域，难以发展出更为通用的智能系统。

为解决这一问题，很多研究人员转而寻求泛化能力更加强大、能够提取数据更一般性的特征表示的模型，而不是为每一个新的任务甚至数据集设计新的模型。很快，人们发现通过迁移学习和多任务学习能够得到这样的模型。在计算机视觉领域，在 ImageNet 这样的大规模图像分类数据集上预训练的模型能够在很多图像分类任务上表现良好；在 NLP 领域，通过在大规模无标注文本上预训练一个语言模型也通常能够给各种下游任务带来很大收益。这些发现表明，共享模型在其他任务上学习到的知识能够显著提升模型的性能，而且通常比改进网络结构所带来的收益更大。

在这一背景下，人们开始思考：是否可以训练单个模型来处理几乎所有的 NLP 任务？近年来，多任务学习被广泛地应用在深层神经网络中，在序列标注、文本分类、机器翻译等多个经典 NLP 任务上都取得了令人鼓舞的效果。随着多任务学习的引入，人们发现很多 NLP 任务可以归纳为统一的模型范式，如问答范式[10]、分类范式[11][12]。同时，越来越多的研究者开始关注模型在多任务上的表现，出现了 decaNLP[10]、GLUE[13] 等大规模多任务评测数据集。可以预见，在预训练模型已经取得了巨大成功的基础上，随着多任务学习算法的成熟以及多任务评测基准的规范化，通用神经网络模型还将给自然语言处理领域乃至人工智能带来更多的惊喜。

1.2 研究进展及现状

本节将简要介绍深度学习背景下的自然语言处理、多任务学习、迁移学习以及它们相结合的研究进展及现状，并阐述了它们之间的联系以及目前存在的不足。

自然语言处理（Natural Language Processing, NLP）是一门旨在使得计算机具

备处理、理解和生成自然语言（人类语言）能力的学科。近年来，以神经网络为代表的深度学习在自然语言处理[6][7][8][9]中取得了广泛的成功。然而，不同于语音、图像等连续实值信号，自然语言是由离散的符号构成，这使其难以直接作为神经网络的输入。为解决这一问题，人们使用低微稠密向量来表征文本的语义信息[14][15]，由于语义信息被分布到向量的各个维度，因此这种方法被称为分布式表示。随着分布式表示的引入，深度学习在自然语言处理领域得到了广泛的应用，卷积神经网络（Convolutional Neural Network, CNN）、循环神经网络（Recurrent Neural Network, RNN）相继被用于处理文本数据，近年来又提出了完全基于自注意力机制的全连接网络 Transformer。这些神经网络的应用使得过去很多难以解决的 NLP 问题上取得了巨大进展。

事实上，文本的分布式表示的好坏对于模型的性能起着至关重要的作用。以分类任务为例，给定数据的一个好的表示，即使简单的线性分类器也能取得非常高的分类准确率[16][17]。进入深度学习时代以来，自然语言处理领域中取得的许多突破都来自于对文本的通用表示方法的研究，如 word2vec[14]，ELMo[18]，BERT[12]等。然而，相较于语音和图像数据，由于文本数据本身的离散性和歧义性，以及标注成本高、难度大等问题，如何得到一个好的文本表示仍然是自然语言处理领域的重大难题。

从机器学习的角度来看，一个好的表示方法除了能够在对应任务上表现良好，还应当具备良好的可迁移性与泛化能力，即能够在相似任务和新数据上获得较好的效果。在自然语言处理领域，常常使用多任务学习（Multi-Task Learning, MTL）和迁移学习（Transfer Learning）的方法来得到这样的文本表示[12][19]。

多任务学习可以追溯到 1993 年[20]，它是指同时使用多个任务对模型进行训练，使其学习到数据的某种泛化表示，该表示能够同时解释这多个任务。在过去的几年里，大量研究人员探索了多任务学习在卷积神经网络和循环神经网络上的应用模式，验证了基于神经网络的多任务学习在文本表示上的有效性。Collobert 等人[19]使用一个简单的卷积网络来同时学习词性标注、语块标注、命名实体识别、语义角色标注、语义相似度和语言模型等多个任务，超越了使用单任务训练

的效果。随着循环神经网络在 NLP 上的广泛应用，研究者开始基于循环网络构造多任务学习框架，在机器翻译[21]、文本分类[22][23]、序列标注[24] 等常见 NLP 任务上均取得了成功。多任务学习的一个关键问题在于如何设计一个高效的共享模式来允许模型共享多个任务的知识。上述提到的工作也大都致力于为所要解决的问题以及采用的网络结构来设计合适的共享模式。

同时，也有大量工作致力于使用迁移学习的范式来获取文本的通用表示，一般做法是利用语言模型[18]][11]、机器翻译[25] 以及其他无监督任务[12] 来预训练一个可迁移的模型。并且，迁移学习与多任务学习本身并不互斥，因此可以同时利用二者的方法，使用多任务预训练迁移模型[26]，也可以在预训练得到的模型的基础上再使用多任务来微调[27][28]。

事实上，多任务学习和迁移学习本质上都是通过共享参数来迁移模型在不同任务中学习到的知识，并以此来提升泛化能力。因此，通常在迁移学习中效果很好的模型也可以应用在多任务学习中。近期，以 Transformer 为预训练网络结构得到的迁移模型 GPT[11] 和 BERT[12] 在诸多自然语言处理任务上取得了极大的提升，这证明了 Transformer 强大的文本表示能力。然而，不同于 CNN 和 RNN，目前还很少有工作研究多任务学习在 Transformer 上的应用模式，已有的少量工作也只是将最传统的多任务共享模式简单地应用在 Transformer 中[27]。

1.3 研究内容

本文试图在一定程度上填补目前多任务学习在 Transformer 结构下的研究空缺。首先，较为系统地考察各种常见多任务学习架构在 Transformer 上的应用效果，然后，根据 Transformer 自身的结构特点设计新的多任务学习架构。

在 CNN 及 RNN 结构中应用多任务学习通常是“纵向”的，即在网络结构的某一层上堆叠任务特定层[24][29]。这种架构蕴含着一个假设：存在某种通用的文本表示，特定的任务表示可以由该通用表示通过简单的变换得到。然而，这样的假设限制了能够同时学习的任务的多样性，难以处理弱相关任务及不同难度的任务。

有一些工作采用了“横向”共享的架构，即允许模型使用其他任务的模型在同一层的隐状态[22][30][?]，然而这些架构也存在两个问题：1）通常需要为每个任务训练一个模型，参数量大且难以扩展；2）各任务模型之间的信息交互难以控制。

而在 Transformer 中，可以很容易地在“横向”以记号的形式进行扩展，并且扩展的记号可以像普通单词一样与句子中的每个单词进行交互。基于这一特性，本文给出了两种新型多任务共享模式，层级-隐式共享模式和层级-显式共享模式，并通过大量实验证明了两种模式的有效性。

1.4 论文结构

本文主要内容包括介绍已有的基于多任务学习的文本表示方法，几种新型的多任务文本表示模型及其实验结果，工作总结及对未来的展望。全文分为五个章节进行介绍，具体结构安排如下：

第 1 章，介绍研究背景及意义，概括本文的研究内容。

第 2 章，介绍相关的理论基础及前沿进展。在 2.1节介绍深度学习的相关概念；在 2.2 节介绍基于神经网络的多任务学习；在 2.3 节介绍深度学习背景下自然语言处理的发展现状；在 2.4 节介绍多任务学习在自然语言处理中的应用。

第 3 章，详细介绍本文研究的模型结构及实现细节。在 3.1 节介绍 Transformer 的模型结构；在 3.2 节介绍几种多任务 Transformer 架构；在 3.3 节给出本文所使用的超参数设置以及代码实现细节。

第 4 章，介绍实验相关的信息。在 4.1 节描述模型应用的具体任务；在 4.2 节给出本文使用的各个数据集的统计信息；在 4.3 节介绍模型在各数据集上的实验结果；最后在 4.4 节对模型进行细粒度的分析并解释其有效性。

第 5 章，对本文的贡献和不足进行总结，回顾相关领域面临的机遇与挑战，并给出了未来的研究方向。

第二章 相关工作

本文研究的内容包含了深度学习、多任务学习与自然语言处理三个主题，这里首先阐述这些主题之间的联系，接着分别介绍它们各自的相关概念与研究进展，最后介绍了三者交叉的一些重要研究工作。

我们利用深度学习与多任务学习来解决自然语言处理中的问题，具体的，即在深层神经网络中使用多任务学习来获得文本的泛化表示。从这一角度来看，深度学习与多任务学习是我们的工具，而自然语言处理为我们的应用场景。同时，深度学习与多任务学习都可以归为机器学习问题，而他们二者又有交叉，如图 2.1 所示。

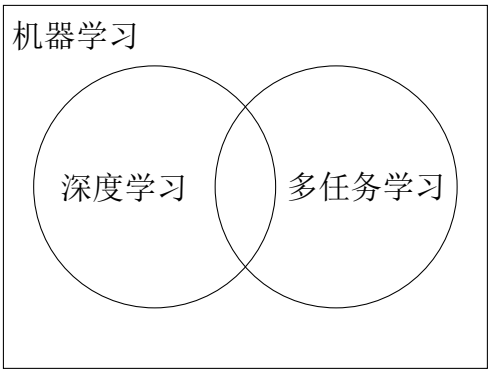

图 2.1　机器学习、深度学习、多任务学习的关系

2.1　深度学习

近年来，深度学习（Deep Learning）发展十分迅速，在人工智能的很多子领域上取得了巨大成功，如语音识别[1][2]、计算机视觉[3][4][5]、自然语言处理[6][7][8][9]等。

从要研究的问题来看，深度学习属于机器学习的一个分支，都是研究如何从有限个样例中通过某种算法总结出一般性规律或模式，并将其应用到新的未知数据上去。例如，通过机器学习或深度学习算法，可以根据历史病例总结出症状与疾病之间的规律，当遇到新的病人时可以利用算法总结出来的规律来帮助判断这

个病人得了什么疾病[1]。

同时，深度学习又与机器学习有很大的不同。从模型的构成来说，深度学习模型一般更加复杂，参数量更多。由于数据从输入到输出需要经过多个线性或非线性组件，信息传递路径较长，因此被称作深度学习。深度学习模型的一个最典型代表就是人工神经网络（Artificial Neural Network，ANN），下面简称神经网络。神经网络是一种受人脑神经系统启发的复杂数学模型，由神经元以及它们之间的连接组成。通过这些神经元的加工，原始数据从底层特征开始经过一道道加工逐渐得到抽象的高层语义特征。

总的来说，神经网络可以被看做一个包含大量参数的复合函数，该函数描述了输入和输出之间的复杂关系，因此可以被用于处理很多模式识别任务，如语音识别、图像识别等。形式化地，神经网络可以这样描述：

$$y = \mathcal{F}(\mathbf{x} \mid \theta). \tag{2-1}$$

其中，函数 $\mathcal{F}$ 受参数 θ 控制。

给定包含大量样例的集合，参数 θ 可以通过随机梯度下降等优化算法得到，求解参数的优化过程就被称为*学习过程*，通过学习得到的参数被称为*可学习参数*，常简称为参数。还有一部分不可以被学习的参数，例如神经网络的层数，被称为*超参数*，意为控制参数的参数，超参数通常根据经验指定或根据实验效果来确定。

前面提到，参数学习需要一个包含大量样例的集合，该集合就是*数据集*。在有监督学习中，每个样例一般包含输入 $\mathbf{x}$ 和输出 y，输出也常被称为标签。其中，输入一般为一个向量，向量的每一个维度表示了样本的一个属性；输出一般为一个标量。包含 n 个样例的集合 $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ 就是数据集。在实际使用时，数据集通常被划分为训练集、验证集（也叫开发集）和测试集。其中，用训练集来让算法学习参数，用验证集来调整算法的超参数，用测试集来衡量模型的优劣。

给定数据集 $\mathcal{D}$ 和模型结构，即 $\mathcal{F}$ 的形式，优化算法可以得到参数 θ 进而确

[1]例子来源：《神经网络与深度学习》，邱锡鹏，https://nndl.github.io

定模型 $\mathcal{F}(\theta)$. 下面介绍一种简单的前馈神经网络结构：多层感知机（Multi-Layer Perceptron, MLP）。一个单隐层的 MLP 可以记作 $\mathcal{F}(\mathbf{x} \mid \theta) = \mathbf{W}_2(f(\mathbf{W}_1\mathbf{x} + b_1)) + b_2$，其中参数 $\theta = \{\mathbf{W}_1, b_1, \mathbf{W}_2, b_2\}$，函数 $f(\cdot)$ 为非线性激活函数，如 ReLU. 多层感知机的结构如图 2.2 所示，为简单起见，图中省略了偏置项 b_1, b_2.

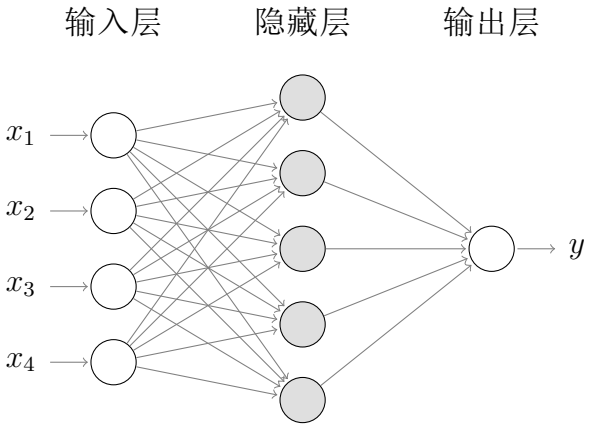

图 2.2　多层感知机

随着深度学习的发展，神经网络的结构也变得日益多样和复杂。上面的多层感知机属于全连接前馈网络，而前馈网络的另一典型代表就是在计算机视觉中被广泛使用的*卷积神经网络*（Convolutional Neural Network, CNN），它与全连接网络的区别在于权重共享和局部连接。除前馈网络外，还有一类网络称为反馈网络，因为反馈网络的神经元可以接收来自本身的反馈信号，从而具备一定的记忆能力，因此也被称为记忆网络。反馈网络的一种典型代表就是*循环神经网络*（Recurrent Neural Network, RNN），它在自然语言处理中得到了广泛应用。另外，近年来图神经网络（Graph Neural Network, GNN）因其在处理图结构数据上的优势也得到了迅速发展。

深度学习能够成功的重要原因在于其强大的特征表示能力。在传统的机器学习中，通常需要人为地构造数据特征，再训练一个学习算法来总结这些构造出来的特征输入与输出之间的关系。例如，在文本分类任务中，传统机器学习的做法一般是利用词袋模型或 TF-IDF 来构造文本特征，再将其作为支持向量机等分类器的输入来进行分类。而在深层神经网络中，算法自动学习原始数据的特征表示，数据从输入到输出的过程 $\mathbf{x} \to \mathcal{F} \to y$ 可以人为地划分为两个阶段：*表示学习*和预

测。所谓表示，就是神经网络中间隐层的状态，上文中 MLP 的中间隐层就可以当作某种浅层表示；而预测，比如分类或回归，一般是在上一层得到的特征表示的基础上进行简单变换得到易于预测的任务特定表示。因此，深度学习又可以看作是一种表示学习，而深度学习的成功也证明了学到好的特征表示的重要意义。

2.2 多任务学习

在机器学习中，我们通常关心模型在某个任务上的某个指标，并通过在某个数据集上训练单个模型或一组集成模型来达成此目标。事实上，有时我们可以同时利用其他数据集甚至其他任务来联合训练模型，从而进一步提升性能，这种方法就是*多任务学习*（Multi-Task Learning, MTL）。

从定义上来说，多任务学习就是一种归纳转移（Inductive Transfer）方法，通过利用包含在相关任务训练信号中的领域特定信息来提升模型的泛化能力[31]。在机器学习中，给定训练集，存在多个假设能够解释训练数据，这些假设构成的空间被称为假设空间，里面的一个假设就对应了一个模型。不同的机器学习算法倾向于选择不同的假设，这种选择偏好被称为归纳偏置（Inductive Bias）。直观上，多任务学习使得算法倾向于选择一种能够同时解释多个任务的假设，也即引入了多个任务的归纳偏置。从表示学习的角度看，这样能够同时适用于多个任务的表示就是泛化的表示。图 2.3 较为直观地给出了多任务学习的一种解释。

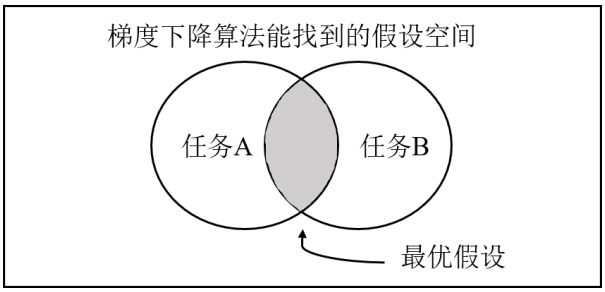

图 2.3 多任务学习的一种直观解释

关于多任务学习为何有效，除了上面的解释（归纳偏置）之外，Caruana 还给出了几种合理的解释[31]：

- **数据增强（Statistical data amplification）** 由于多个任务的数据集通常是不

同的，使用多个相关任务对同一个模型进行训练相当于增大了训练数据量。

- **特征选择（Attribute selection）** 如果任务噪声较大，或者数据高维而数据量有限，模型很难分辨相关特征和无关特征，而多任务学习可以帮助模型选出相关特征，因为这些特征通常在多个任务中是共用的。也就是说，其他任务为模型选择特征提供了额外的证据。
- **窃听（Eavesdropping）** 存在某些特征对于任务 A 易于学习，而对于任务 B 则难以学习。通过多任务学习，模型可以在执行任务 B 时使用任务 A 学到的特征。

在深度学习之前，多任务学习已经被应用在线性模型、核方法、决策树、多层感知机等传统机器学习算法上，有大量的研究集中在稀疏正则化[32][33] 以及学习任务之间的关系[34][35] 上。

随着深度学习的发展，多任务学习开始应用在深层神经网络中，并在自然语言处理[19]、计算机视觉[30]、语音识别[36]、药物设计[37] 等众多应用场景中取得了成功。同时，多任务学习作为一种模型无关的方法，在卷积神经网络[19][30]、循环神经网络[22]、图网络[38] 上都可以应用。

然而，无论是在传统机器学习算法上还是在深层神经网络上，多任务学习的核心观点都在于知识共享。如何为特定任务和学习算法设计合适的共享模式一直是多任务学习的重要问题。从多任务学习的应用方式来看，可以大概分为下面四种共享模式[39]：

- **硬共享模式：** 让不同任务的模型共享一些公用模块（比如神经网络的低层）来提取一些通用特征表示，然后再针对每个不同的任务设置一些私有模块（比如神经网络的高层）来提取任务特定的特征表示。
- **软共享模式：** 不显式地设置共享模块，但每个任务都可以“窃取”其他任务的模型学习到的特征表示。窃取的方式多种多样，比如可以直接使用其它任务的隐状态，也可以使用注意力机制或门控机制来选取有用的特征。
- **分层共享模式：** 一般神经网络中不同层抽取的特征类型不同。底层一般抽取一些低级的局部特征，高层抽取一些高级的抽象语义特征。因此如果多任务

学习中不同任务也有级别高低之分，那么一个合理的共享模式是让低级任务在底层输出，高级任务在高层输出。

- **共享-私有模式：**一个更加分工明确的方式是将共享模块和任务特定（私有）模块的职责分开。共享模块捕捉一些跨任务的共享特征，而私有模块只捕捉和特定任务相关的特征。最终的表示由共享特征和私有特征共同构成。

四种共享模式如图 2.4 所示。

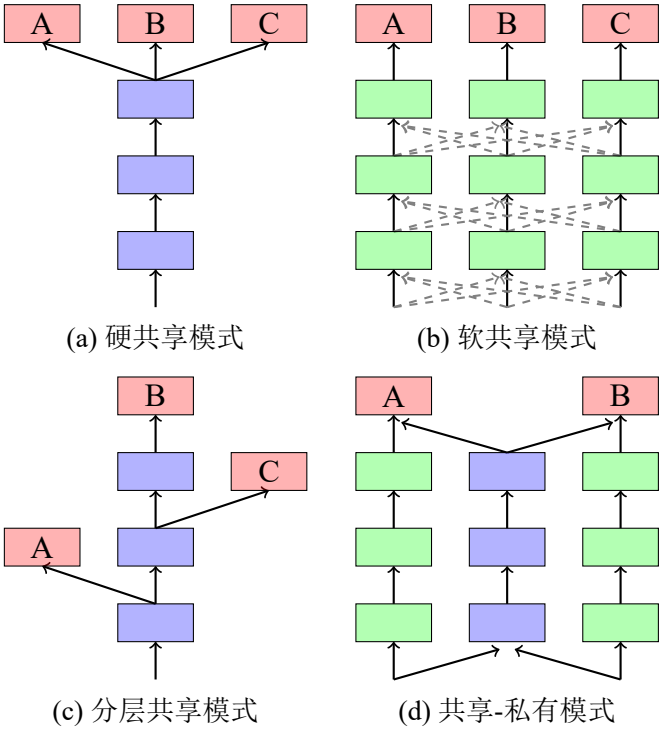

图 2.4　多任务学习的几种常见共享模式

2.3　自然语言处理

2.4　多任务自然语言处理

第三章 模型

3.1　Transformer

3.2　多任务 Transformer

3.2.1 S-P 结构

3.2.2 S-C 结构

3.2.3 L-I 结构

3.2.4 L-E 结构

3.3　实现细节

第四章 实验

4.1　任务描述

4.2　数据集

4.3　实验结果

4.4　实验分析

第五章 总结与展望

致谢

在此论文完成之际，... 谨向... 的人表示由衷的感谢！

首先，我要衷心感谢我的指导老师...

...

感谢齐飞副教授提供的 XDU-*THESIS* 模版，它的存在让我的论文写作轻松自在了许多，让我的论文格式规整漂亮了许多。

参考文献

[1] MIKOLOV T, DEORAS A, POVEY D, et al. Strategies for training large scale neural network language models[C/OL] // NAHAMOO D, PICHENY M. 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011, Waikoloa, HI, USA, December 11-15, 2011. [S.l.] : IEEE, 2011 : 196 – 201.
https://doi.org/10.1109/ASRU.2011.6163930.

[2] LI X, HONG C, YANG Y, et al. Deep neural networks for syllable based acoustic modeling in Chinese speech recognition[C/OL] // Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013, Kaohsiung, Taiwan, October 29 - November 1, 2013. [S.l.] : IEEE, 2013 : 1 – 4.
https://doi.org/10.1109/APSIPA.2013.6694176.

[3] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J/OL]. Commun. ACM, 2017, 60(6) : 84 – 90.
http://doi.acm.org/10.1145/3065386.

[4] FARABET C, COUPRIE C, NAJMAN L, et al. Learning Hierarchical Features for Scene Labeling[J/OL]. IEEE Trans. Pattern Anal. Mach. Intell., 2013, 35(8) : 1915 – 1929.
https://doi.org/10.1109/TPAMI.2012.231.

[5] TOMPSON J J, JAIN A, LECUN Y, et al. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation[C/OL] // GHAHRAMANI Z, WELLING M, CORTES C, et al. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. 2014 : 1799 – 1807.
http://papers.nips.cc/paper/5573-joint-training-of-a-convolutional-netw

[6] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural Language Processing (Almost) from Scratch[J/OL]. Journal of Machine Learning Research, 2011, 12 : 2493 – 2537.
http://dl.acm.org/citation.cfm?id=2078186.

[7] BORDES A, CHOPRA S, WESTON J. Question Answering with Subgraph Embeddings[C/OL] // MOSCHITTI A, PANG B, DAELEMANS W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. [S.l.]: ACL, 2014: 615 – 620.
http://aclweb.org/anthology/D/D14/D14-1067.pdf.

[8] JEAN S, CHO K, MEMISEVIC R, et al. On Using Very Large Target Vocabulary for Neural Machine Translation[C/OL] // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. [S.l.]: The Association for Computer Linguistics, 2015: 1 – 10.
http://aclweb.org/anthology/P/P15/P15-1001.pdf.

[9] SUTSKEVER I, VINYALS O, LE Q V. Sequence to Sequence Learning with Neural Networks[C/OL] // GHAHRAMANI Z, WELLING M, CORTES C, et al. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. 2014: 3104 – 3112.
http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-netwo

[10] MCCANN B, KESKAR N S, XIONG C, et al. The natural language decathlon: Multitask learning as question answering[J]. arXiv preprint arXiv:1806.08730, 2018.

[11] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf, 2018.

[12] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[13] WANG A, SINGH A, MICHAEL J, et al. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding[C/OL] // LINZEN T, CHRUPALA G, ALISHAHI A. Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018. [S.l.]: Association for Com-

putational Linguistics, 2018 : 353 – 355.
https://aclanthology.info/papers/W18-5446/w18-5446.

[14] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and their Compositionality[C/OL] // BURGES C J C, BOTTOU L, GHAHRAMANI Z, et al. Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.. 2013 : 3111 – 3119.
http://papers.nips.cc/paper/5021-distributed-representations-of-words-a

[15] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global Vectors for Word Representation[C/OL] // MOSCHITTI A, PANG B, DAELEMANS W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. [S.l.] : ACL, 2014 : 1532 – 1543.
http://aclweb.org/anthology/D/D14/D14-1162.pdf.

[16] TENNEY I, XIA P, CHEN B, et al. What do you learn from context? Probing for sentence structure in contextualized word representations[J], 2018.

[17] LIU N F, GARDNER M, BELINKOV Y, et al. Linguistic Knowledge and Transferability of Contextual Representations[J]. arXiv preprint arXiv:1903.08855, 2019.

[18] PETERS M E, NEUMANN M, IYYER M, et al. Deep Contextualized Word Representations[C/OL] // WALKER M A, JI H, STENT A. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers). [S.l.] : Association for Computational Linguistics, 2018 : 2227 – 2237.
https://aclanthology.info/papers/N18-1202/n18-1202.

[19] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning[C/OL] // COHEN W W, MCCALLUM A, ROWEIS S T. ACM International Conference Proceeding Series, Vol 307 : Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008. [S.l.] : ACM, 2008 : 160 – 167.
https://doi.org/10.1145/1390156.1390177.

[20] CARUANA R. Multitask Learning: A Knowledge-Based Source of Inductive Bias[C] // Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June 27-29, 1993. [S.l.]: Morgan Kaufmann, 1993: 41 – 48.

[21] DONG D, WU H, HE W, et al. Multi-Task Learning for Multiple Language Translation[C/OL] // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. [S.l.]: The Association for Computer Linguistics, 2015: 1723 – 1732. http://aclweb.org/anthology/P/P15/P15-1166.pdf.

[22] LIU P, QIU X, HUANG X. Recurrent Neural Network for Text Classification with Multi-Task Learning[C/OL] // KAMBHAMPATI S. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016. [S.l.]: IJCAI/AAAI Press, 2016: 2873 – 2879. http://www.ijcai.org/Abstract/16/408.

[23] LIU P, QIU X, HUANG X. Adversarial Multi-task Learning for Text Classification[C/OL] // BARZILAY R, KAN M. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. [S.l.]: Association for Computational Linguistics, 2017: 1 – 10. https://doi.org/10.18653/v1/P17-1001.

[24] SØGAARD A, GOLDBERG Y. Deep multi-task learning with low level tasks supervised at lower layers[C/OL] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers. [S.l.]: The Association for Computer Linguistics, 2016. http://aclweb.org/anthology/P/P16/P16-2038.pdf.

[25] MCCANN B, BRADBURY J, XIONG C, et al. Learned in Translation: Contextualized Word Vectors[C/OL] // GUYON I, von LUXBURG U, BENGIO S, et al. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. 2017: 6297 – 6308. http://papers.nips.cc/paper/7209-learned-in-translation-contextualized-word-vect

[26] SUBRAMANIAN S, TRISCHLER A, BENGIO Y, et al. Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning[C/OL] // 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. 2018.
https://openreview.net/forum?id=B18WgG-CZ.

[27] LIU X, HE P, CHEN W, et al. Multi-Task Deep Neural Networks for Natural Language Understanding[J]. arXiv preprint arXiv:1901.11504, 2019.

[28] ANONYMOUS. BAM! Born-Again Multi-Task Networks for Natural Language Understanding[H]. 2018.

[29] ZHENG R, CHEN J, QIU X. Same Representation, Different Attentions: Shareable Sentence Representation Learning from Multiple Tasks[C/OL] // LANG J. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.. [S.l.] : ijcai.org, 2018 : 4616 – 4622.
https://doi.org/10.24963/ijcai.2018/642.

[30] MISRA I, SHRIVASTAVA A, GUPTA A, et al. Cross-Stitch Networks for Multi-task Learning[C/OL] // 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. [S.l.] : IEEE Computer Society, 2016 : 3994 – 4003.
https://doi.org/10.1109/CVPR.2016.433.

[31] CARUANA R. Multitask Learning[J/OL]. Machine Learning, 1997, 28(1) : 41 – 75.
https://doi.org/10.1023/A:1007379606734.

[32] ARGYRIOU A, EVGENIOU T, PONTIL M. Multi-Task Feature Learning[C/OL] // SCHÖLKOPF B, PLATT J C, HOFMANN T. Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006. [S.l.] : MIT Press, 2006 : 41 – 48.
http://papers.nips.cc/paper/3143-multi-task-feature-learning.

[33] LOUNICI K, PONTIL M, TSYBAKOV A B, et al. Taking Advantage of Sparsity in Multi-Task Learning[C/OL] // COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009. 2009.
http://www.cs.mcgill.ca/%7Ecolt2009/papers/008.pdf#page=1.

[34] EVGENIOU T, MICCHELLI C A, PONTIL M. Learning Multiple Tasks with Kernel Methods[J/OL]. Journal of Machine Learning Research, 2005, 6 : 615 – 637.
http://jmlr.org/papers/v6/evgeniou05a.html.

[35] JACOB L, BACH F R, VERT J. Clustered Multi-Task Learning: A Convex Formulation[C/OL] // KOLLER D, SCHUURMANS D, BENGIO Y, et al. Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008. [S.l.] : Curran Associates, Inc., 2008 : 745 – 752.
http://papers.nips.cc/paper/3499-clustered-multi-task-learning-a-convex-formulat

[36] DENG L, HINTON G E, KINGSBURY B. New types of deep neural network learning for speech recognition and related applications: an overview[C/OL] // IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013. [S.l.] : IEEE, 2013 : 8599 – 8603.
https://doi.org/10.1109/ICASSP.2013.6639344.

[37] RAMSUNDAR B, KEARNES S M, RILEY P, et al. Massively Multitask Networks for Drug Discovery[J/OL]. CoRR, 2015, abs/1502.02072.
http://arxiv.org/abs/1502.02072.

[38] LIU P, FU J, DONG Y, et al. Multi-task Learning over Graph Structures[J]. arXiv preprint arXiv:1811.10211, 2018.

[39] 邱锡鹏. 神经网络与深度学习 [M/OL]. 2019.
https://nndl.github.io/.

[40] GHAHRAMANI Z, WELLING M, CORTES C, et al. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada[C/OL]. 2014.
http://papers.nips.cc/book/advances-in-neural-information-processing-systems-27-

[41] MOSCHITTI A, PANG B, DAELEMANS W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL[C]. [S.l.]: ACL, 2014.

[42] ANON. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers[C/OL]. [S.l.]: The Association for Computer Linguistics, 2015.
http://aclweb.org/anthology/P/P15/.