文章编号：2095-2163(2020)06-0001-04 中图分类号：TP18 文献标志码：A

基于迁移学习的中文阅读理解

孙一博，秦 兵
（哈尔滨工业大学 计算机科学与技术学院社会计算与信息检索研究中心，哈尔滨 150001）

摘 要：深度学习在解决自然语言处理中的机器阅读理解任务方面引起了越来越多的关注。尽管已经取得了巨大的成功，但典型的深度学习方法依赖大量的标记数据，在解决许多实际的问题中通常是不可用的。最近，针对这种情况探索了一种知识转移范式，其目的是利用源问题域的丰富训练数据，帮助模型在目标域中更有效地解决问题。本文重点研究基于渐进神经网络用迁移学习的方式解决中文机器阅读理解任务，提出了一种渐进式学习模型，该模型通过使用交叉注意适配器改进渐进神经网络，进而使其具有在异构输入之间传递知识的能力。通过在中国高考阅读理解数据集进行实验，结果表明本文所提出的模型优于所有的基线模型。

关键词：深度学习；机器阅读理解；迁移学习

Transfer Learning for Chinese Machine Reading Comprehension

SUN Yibo, QIN Bing
(Information Retrieval Lab, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Deep learning (DL) has attracted increasing attentions on solving the task of machine comprehension (MC) in natural language processing (NLP). Nevertheless, despite the remarkable success it has exhibited, typical DL methods demands a large amount of labeled data to be effective, which is often unavailable in many practical problem domains, such as Chinese reading comprehension. Recently, studies towards this situation have explored a knowledge transfer paradigm, which aims at leveraging highly heterogeneous data from the well-trained source problem domains for more effective problem-solving in the target ones. Taking this cue, this paper also contributes to this paradigm and our special interest lies in solving the Chinese MC task by improving a recent proposed progressive neural network (PNN). Specifically, we propose a progressive learning model which facilitates to complement PNN with a Cross Attention Adapter. As a result, the model endows PNN with the capability to transfer knowledge between heterogeneous inputs. To verify the efficacy of the proposed transfer paradigm, comprehensive experiments are investigated on a Chinese college entrance examination dataset. The empirical results have shown that the proposed model outperforms Progressive Neural Network with conventional adapter and all other baselines settings.

【Key words】 Deep Learning; Machine Comprehension; Transfer Learning

0 引 言

近年来，机器阅读理解任务引起了学界越来越多的关注，其旨在验证机器以类似于人的方式理解文档的能力。基于神经网络的模型在机器阅读理解任务上大获成功，但大多数都是针对大量数据的数据集而设计的，例如：CNN，Squad，NewsQA 等。本文要研究的中文高考阅读理解任务，由于其数据量较小，主流的神经网络模型在其上容易收敛失败。

为了解决数据稀缺的问题，本文采用基于迁移学习的方法，即使用来自数据丰富的源领域的知识来帮助模型在数据稀少且不宜获得的目标领域进行高效的训练[1]。迁移学习在神经网络模型中的一个比较直观的应用是微调（fine-tuning），即用源领域的数据对模型进行预训练，应用训练得到权重来初始化模型，随后在目标领域的数据上通过反向传播对模型的权重进行微调。

本文基于渐进神经网络（Progressive neural network）进行中文机器阅读理解的研究。渐进神经网络是近期提出的一种迁移学习框架，它使用横向的链接机制来保留从历史任务中学习得到的权重，从而在一定程度上解决了迁移学习中的灾难性遗忘问题[2]。本文针对机器阅读理解任务对渐进神经网络进行了扩展，使用注意力机制对不同任务间的表示进行计算，并且使用长短记忆网络对得到的表示建模。利用中国高考历届试卷构建中文阅读理解数据集 CTARC 作为目标数据集，并使用已有的英文阅读理解数据集 RACE 作为源数据集。实验结果表明，本文提出的方法优于所有的基线模型。

作者简介：孙一博（1990-），男，博士研究生，主要研究方向：自然语言处理、深度学习；秦 兵（1968-），女，博士，教授，主要研究方向：自然语言处理、情感分析、知识图谱等。

收稿日期：2020-04-12

1 方法

1.1 渐进神经网络

如图 1 所示,一个渐进神经网络由若干个互相链接的列组成,其中每列代表一个由参数 $\Theta^{(j)}$ 构成的 L 层的神经网络,每层都包含隐层输出 $h_i^{(j)} \in \mathbb{R}^{n_i}$, 其中 n_i 是第 j 列第 i 层的输出维度。在第一列中,代表建模第一个任务的神经模型的训练过程结束后,可以得到其参数 $\Theta^{(1)}$ 。 随后,在两个任务的情况下,当第二个任务来临时,保持在第一个任务中学习到的参数不变,并使用随机初始化的新参数 $\Theta^{(2)}$ 初始化代表第二个任务的神经网络模型,其中该模型第 i 层的输入包含两个部分:一部分来自第二个任务自身的输入 $h_{i-1}^{(2)}$,另一部分来第一个任务上一层输出 $h_{i-1}^{(1)}$ 的横向连接。具体的,当推广到 K 个任务,这种用于融入历史任务信息的横向链接可以形式化为(1)式:

$$h_i^{(k)} = f\left(W_i^{(k)} h_{i-1}^{(k)} + \sum_{j<k} U_i^{(k:j)} h_{i-1}^{(j)}\right). \tag{1}$$

其中, $W_i^{(k)} \in R^{n_i \times n_{i-1}}$ 是第 k 列第 i 层的参数矩阵,$U_i^{(k:j)} \in R^{n_i \times n_j}$ 是从第 j 列的第 $i-1$ 层到第 k 列的第 i 层的横向链接矩阵。h_0 是网络的输入。f 是任意的激活函数,如:Relu 或者 Sigmoid 函数。

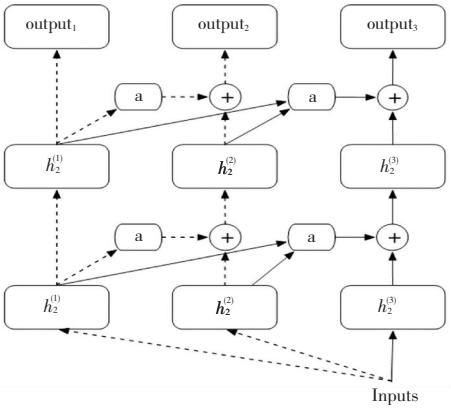

图 1 渐进神经网络

Fig. 1 Progressive Neural Network

1.2 适配器

上文中的横向连接 $U_i^{(k:j)} \in R^{n_i \times n_j}$ 可以 被扩展到非线性函数,将这种非线性的横向链接称为适配器。在适配器的模式下,首先将历史模型乘以一个标量,然后将其收尾相连,并送入一个具有非线性激活函数的前馈神经网络。采用这种方式,上述标量可以决定历史中各个任务中学习到的特征的重要性,同时在前馈神经网络中的投影操作也可以解决任务数量过多后参数维度膨胀的问题。

形式化的,用 $h_{i-1}^{(<k)} = [h_{i-1}^{(1)} \cdots h_{i-1}^{(j)} \cdots h_{i-1}^{(k-1)}]$ 表示从第 1 列到第 $k-1$ 列第 $i-1$ 层维度为 $n_{i-1}^{(<k)}$ 的历史输出的收尾相连,用 $\alpha_{i-1}^{(<k)}$ 表示每个前序输出对应的可学习的标量。在省略偏置项的条件下,渐进神经网络中用适配器模式表达的第 k 列中的第 i 层的输入如公式(2) 所示:

$$\boldsymbol{h}_i^{(k)} = \boldsymbol{\sigma}(\boldsymbol{W}_i^{(k)} \boldsymbol{h}_{i-1}^{(k)} + \boldsymbol{U}_i^{(k:j)} \boldsymbol{\sigma}(\boldsymbol{V}_i^{(k:j)} \boldsymbol{\alpha}_{i-1}^{(<k)} \boldsymbol{h}_{i-1}^{(<k)})). \tag{2}$$

其中, $V_i^{(k:j)} \in R^{n_{i-1} \times n_i^{(<k)}}$ 是投影矩阵。

1.3 交叉注意力适配器

在渐进神经网络架构中, 单列的所有层都可以看作是一个单变量函数 $y=f(x)$,这种架构适用于强化学习,其中神经模型的设计是为了解决特定的马尔可夫决策过程。当尝试将渐进神经网络框架应用于阅读理解模型时,为了找到文档中与问题相关的部分,注意力机制几乎在所有的为了阅读理解设计的神经模型中都经常被使用。注意力机制层通常以文档表示和句子表示作为输入,根据问题输出文档表示的加权求和,使得该层成为一个二元函数,不符合原来的渐进神经网络框架。

为了克服这个限制,一个直接的方法是使用公式(2)分别处理两个输入,而不考虑它们之间的交互作用,这使得当前任务中的注意力层的输入和前一个任务中的输入属于不同的语义空间。使用两个不同的语义空间的组合可能会破坏原有的意义,这种方式很难从前面的注意层中学习到注意力权重。

如图 2 所示,为了能更好地学习注意力知识,本文提出了一种方法,旨在通过让两个语义空间的两部分跨越不同的任务来进行注意力机制,缩短两个语义空间之间的距离。形式化的,用二元组 (l_{i-1}^j, r_{i-1}^j) 表示第 j 列第 $i-1$ 层的两个输出,用 g 表示注意力机制的二元函数,每个函数 g 都有不同的参数。在当前列 k,通过将 l_{i-1}^j, r_{i-1}^j 对前序相应位置做注意力机制,可以得到如下的两个列表(3)和(4):

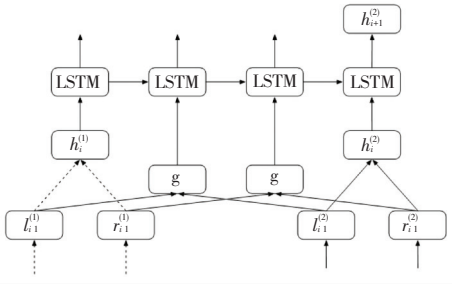

图 2 交叉注意力适配器

Fig. 2 Cross Attention Adapter

$$A = [g(l_{i-1}^1, r_{i-1}^k), \cdots, g(l_{i-1}^j, r_{i-1}^k), \cdots, g(l_{i-1}^{k-1}, r_{i-1}^k)], \tag{3}$$

$$B = [g(l_{i-1}^k, r_{i-1}^1), \cdots, g(l_{i-1}^k, r_{i-1}^j), \cdots, g(l_{i-1}^k, r_{i-1}^{k-1})]. \tag{4}$$

随后也可以得到前序二元组之前的内部注意力结果(5)。

$$C = [g(l_{i-1}^1, r_{i-1}^1), \cdots, g(l_{i-1}^j, r_{i-1}^j), \cdots, g(l_{i-1}^{k-1}, r_{i-1}^{k-1})]. \tag{5}$$

实际上，列表 C 代表不使用外部链接的条件下计算第 i 层的注意力机制。在跳过代表注意力机制的第 i 层后，可以用渐进神经网络的方式计算第 $i+1$ 层的输入 $h_{i+1}^{(k)}$。为了更好的获得组合语义性的表达，相对于使用带有非线性的前馈神经网络，本文使用循环神经网络来构建适配器模块：首先，用列表 A, B, C 和当前的内部注意力结果构建了序列 $T = [C, B, A, g(l_{i-1}^k, r_{i-1}^k)]$。用长短记忆神经网络(LSTM)按(6)、(7)式得到其组合语义表示：

$$n_i = \mathrm{LSTM}(n_{i-1}, t_i), i = 1, \cdots, 3K, \tag{6}$$

$$h_{i+1}^{(k)} = n_{3K}. \tag{7}$$

1.4 Stanford AR 模型

Stanford Attentive Reader cite 是一个强大的模型，在 CNN 数据集上取得了最先进的结果。作者声称，该模型在这个数据集上的性能几乎达到了上限。

用 (p, q, o) 来代表由文章，问题，选项，构成三元组。$p = p_1, \cdots, p_m$ 和 $q = q_1, \cdots, q_l$ 是长度为 m 和 l 的文章词序列和问题词序列，$o = o_1, \cdots, o_C$ 是包含 C 个候选答案的答案集合，每个答案也是一个句子。任务的目标就是在集合 o 中找到正确的答案。

本文首先用双向 GRU 将 p 和 q 表示为 $\widetilde{\boldsymbol{p}_1}, \widetilde{\boldsymbol{p}_2}, \cdots, \widetilde{\boldsymbol{p}_n}$ 和 $\boldsymbol{q}$。随后用注意力机制将文章中与问题最相关的部分总结为向量 t。根据的设置，使用双线性的注意力形式[3]：

$$\alpha_i = softmax_i \boldsymbol{q}_T \boldsymbol{W}_s \widetilde{\boldsymbol{p}_i}, \tag{8}$$

$$\boldsymbol{t} = \sum\nolimits_i \alpha_i \widetilde{\boldsymbol{p}_i}. \tag{9}$$

类似的，使用双向 GRU 将选项 o_i 编码为 $\boldsymbol{o}_i$，使用双线性注意力机制计算选项与文章 t 的相关性得分，并将该得分送入 softmax 函数来得到最终的概率分布。具体的，第 i 个选项是正确答案的概率计算公式(10)为：

$$p_i = softmax_i \boldsymbol{t}^{\mathrm{T}} \boldsymbol{W}_t \boldsymbol{o}_i. \tag{10}$$

2 实　验

2.1 数据集

RACE 数据集是从中国三家大型免费的公共网站上收集到的原始数据，其中阅读理解题都是从中国老师设计的英语考试中提取出来的。清洗前的数据共计 137 918 篇文章、519 878 道题，其中初中组有 38 159 篇文章、156 782 道题，高中组有 99 759 篇文章、363 096 道题。清洗后的数据量如表 1 所示。

表 1　数据集划分

Tab. 1　The separation of Dataset

训练集子集	RACE				CTARC			
	Train	Dev	Test	All	Train	Dev	Test	All
文章数量	25 137	1 389	1 407	27 933	136	20	20	176
问题数量	87 866	4 887	4 934	97 687	390	42	54	486

本文收集了 2004～2016 年中国高考的原始资料，其中阅读理解题提取了中国高考专家设计的科普文章阅读理解部分的阅读理解题。该部分中的文章主要是以科普类文章为主。为了对原始数据进行清洗，进行了以下筛选步骤。首先，删除所有与问题设置格式不一致的问题和试题，例如：如果某道题的选项数不是 4 个，就会被删除。同时，还删除了所有含有“下划线”或“段落”等关键词的问题，因为下划线和段落段信息的效果难以重现。经过过滤步骤后，一共得到 176 个段落和 486 个问题。为了进行数据增强，用谷歌翻译将得到的数据进行英文翻译，得到了 CTARC-E 数据集。CTARC 数据集的详细统计见表 2。

表 2　数据集统计

Tab. 2　Dataset statistis

数据集	RACE	CTARC	CTARC-E
文章长度	321.9	1634.68	3735.24
问题长度	10.0	36.95	93.04
选项长度	5.3	60.88	148.25
词表大小	136 629	13 936	9 505

2.2 实验结果

实验结果如表 3 所示，其中 RACE single 代表在 RACE 数据集上训练 Stanford AR 模型，CTARC single in Chinese 代表在 CTARC 数据集上训练 Stanford AR 模型，CTARC single in English 代表在

(下转第 11 页)