基于记忆网络的阅读理解回答技术

Darshan Kapashi
Department of Computer Science
Stanford University
Stanford, CA 94305
darshank@stanford.edu

Pararth Shah
Department of Computer Science
Stanford University
Stanford, CA 94305
darshank@stanford.edu

摘要

在这项工作中，我们将调研利用深度神经网络外加一个记忆组件去构建一个问答系统的任务。我们的目标是使用记忆神经网络和（在[10]和[8]中被描述的）它的扩展并将其应用到（[9]所介绍的）bAbI 问答任务。和例如 bAbI 这样的仿真数据集不同的是，普通的记忆神经网络系统还不足以在像维基问答和MC测试这样的真实问答数据集上达到令人满意的性能。我们将探索已被提出的记忆神经网络系统的扩展，去使得它在这些复杂的数据集上运作。

1. 导论

自然语言处理长期以来的目标是发展一个能够与人类参与者进行自然语言对话的具有普遍目的的人工智能模型。但是，自动评估一个模型在一般性的对话中表现出的性能是一件困难的事情，这也使得人们很难设计一个学习方法去提高模型的性能。但是问答任务可以容易地满足这个标准，因为模型对一个问题的回答可以和预期的答案进行比对。此外，问答任务的适用范围是很广泛的，正如很多自然语言处理任务都能够重构成问答任务。这预示着，在问答任务的基础上设计一个更好的模型去提高模型的正确率和效率是非常有效的。

从根本上而言，问答系统的两个主要任务是：检索和推断。问答系统必须存储一些知识，在一些方便的内在描述和随后的内容，他必须通过搜索这个知识库并推断能够帮助到回答向系统提出的问题的一些信息。在这个处理过程中，对问题文本的分析需要满足推断的特定形式，某种能够经由明确的逻辑规则完成的形式，比如基于微积分去推断被问及的事物、经由诸如 RNN 或者 LSTM 等一系列预测器的训练后的网络参数。这两个任务（检索与推断）对于问答系统的成功是非常

关键的。

在这项工作中，我们将介绍旨在提高对特定问答任务的性能的记忆网络框架的简单扩展。下一个章节包括了在这个领域相关工作的简单描述。第三章节包括了涉及到的问答任务的正式描述。第四章节提供了一个记忆网络各个组件的简要概览和我们对其进行的扩展的细节。第五章节讲呈现一些我们实验的结果，并且最后在第六章节中我将将讨论结论和未来的工作。

2. 背景

显而易见的是，人工智能和机器学习的现代实现方法都极度依赖于一定范围的方法。早期的人工智能使用规则和本体论的观点，但是这些都是启发式的、脆弱的和不可扩展的。即使机器学习的方法已经在通过很多方法扩展到处理结构化的数据，主要的基础方法基本总体上是统计的：机器学习的监督设定通常考虑到一族模型、一个损耗函数和一个带标签的数据集，并且其目标是寻找在对范例之外数据的计算错误达到最小的模型。

但是，直到现在问答系统的两个主要工作都还没有在通过大规模受监督的记忆网络和强大的推断模型的结合取得成功。（[2]展示的）菲德尔的工作是传统问答系统的代表，它们是基于机器学习的目标函数的，也就是通过查询一个从网络中抽取的数据库，这个知识库和提问文本中的需要被考虑到的词汇和句法的模式都已被人工标记出。相对这种技术而言，我们最近已经看到使用诸如 RNNs 和 LSTMs 等记忆单元的深度神经网络是一系列强大的预测器，它们能够被高效地训练而学会去通过文本中一系列词汇间的依赖关系做出推断。它们的不足在于缺乏一种结构性的记忆成分，这种记忆成分可以通过允许网络在回答问题时只专注于相关的信息来简化推断任务。

记忆网络（也就是[10]中威斯顿的工作）是近期提出的模型，旨在学习如何把握推断部分与一个长期记忆部分的因果关系。其记忆的工作模式就像是一个能够从过去唤起一些事实的知识库。把深度网络和一个为了其他任务的记忆组件相结合的类似方法已经在最近被推出了。尤其是对于问答任务，这个模型试图去学习一个估值函数对相关的记忆进行排序。在预测的时候，这个模型根据估值函数找到 k 个相关记忆并基于这些相关记忆决定它的输出。本课题所希望的是，即使 LSTM 在一个复杂的问答任务中或许表现得很差，但是由相关记忆决定的 LSTMs 将会有效很多。

3. 问题陈述

在很大程度上，这项工作的目标是构建一个能够回答由人类用自然语言提出的问题的模型。这一任务包括阅读一个文本，或许是编成成一个故事的几个句子或者组成一个知识库的一系列事实。然后，某个问题会基于已给出的文本被提出来。这个模型被期望去阅读问题并输出一个答案，或许是一个单词或者一个自然语言的句子（我们将两者视作独立的任务）。

Mary moved to the bathroom. John went to the hallway. Daniel went back to the hallway. Sandra moved to the garden. Q: Where is Mary? A: bathroom Q: Where is Daniel? A: hallway

图一：由四个句子的文本和两个基于文本的问题组成的一个问答任务样例

在图一中展示了一个示范任务。这个文本从一个由一些人物、对象和地点组成的模式中生成的。在这个任务中的问题是属于由一个简单的事实支持的基本概要类型的，而后答案依据了文本中的一个简单句子提供的信息。

([9]中展示的)威斯顿的工作已经展示了总共二十个不同难度水平的问答任务，每一个任务测试了人工智能模型不同的记忆和推断能力。为了覆盖一个广泛的语言理解能力的集合，例如事实型问答、否定、计数、同义词、连接词、归纳、推理、位置推理、路径查找等等，这些任务是专门设计的。我们将不在这里从论文模仿所有这些任务，但是我们鼓励读者去原文中阅读具体的描述。

3.1 数据集

3.1.1 bAbI 问答任务

[9]所描述的每一个任务的数据都是从仿真中产生的，类似于上面的例子。作者们分享了一个针对 20 个任务[1]的标准化数据集，其中每项任务有 1000 个训练问题和 1000 个测试问题。它作为一个标准基准被呈现，用于测试基于记忆的问答模型的性能差异。

3.1.2 MC 测试

经作者们描述，MC 测试是一个由故事和相关问题组成的用于对机器理解文本进行研究的开源数据集。每一个故事都是一系列伴随着有四个选项的问题的句子。

在图二中展示了一个示范故事。

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home. Q: Where did James go after he went to the grocery store? (A) his deck (B) his freezer (C) a fast food restaurant (D) his room

图二：由一篇文本和基于文本的多选问题组成的一个 MC 测试理解问题样例

MC 测试数据集在以下几个方面和 bAbI 问答数据集区分来开：

1. MC 测试是一个开域，但是被限制于能被七岁孩子理解的概念

2. MC 使用了真实的自然语言写作的小说故事，而不像 bAbI 那样使用仿真数据

3. 故事最多有 50 个句子而且单词量有大约 2000 个（在 MC160 中）和大约 4000 个（在 MC500 中）。相比之下，bAbI 有大约 15 个句子并且单词量有大约 40 个

3.1.3 维基问答

这个数据集有一个维基文章的全集，并且手工生成了基于事实的问题和回答。他的全集包括真实的维基文章，每一篇文章有大致 100 到 2000 个句子和 40000 多个词的单词量。事实型问题都是开放性答案的并且没有多个选项，同时在这个任务重的语言复杂性比 bAbI 或者 MC 测试都要高出很多，因此毫不意外的是在这个数据集上表现得很好要困难很多。

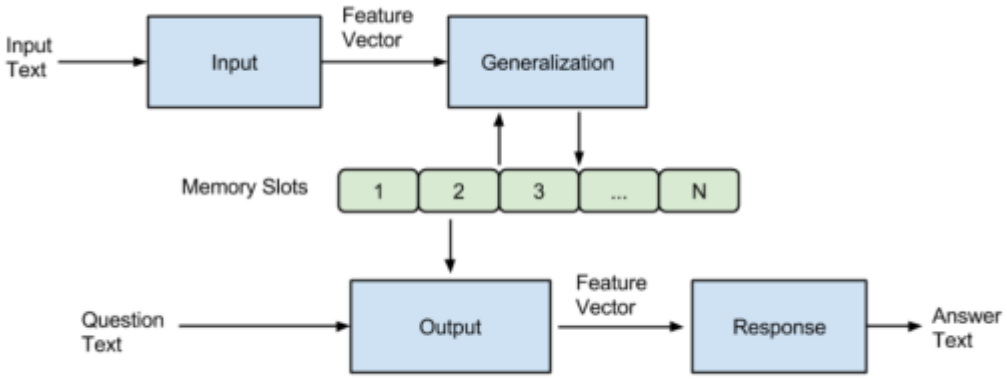
4. 技术方法

4.1 受监督的记忆神经网络

受监督的记忆神经网络（或者记忆神经网络）是被设计去解决如何把一个 LSTM 和记忆组件相结合以实现推断这一确切问题的。我们将依据[10]所描述的模型并引导读者去阅读原文以获取记忆网络的具体描述。作为一个间接的概览，我们在图三中呈现了记忆神经网络的结构。记忆神经网络被分为四个部分:

1. **输入**，将传入的输入文本转换成内部的特征表达

2. **生成**，同时处理输入特征向量和当前的记忆，并决定用哪个时段存储新信息。它还能根据新信息修改/删除任何更早期的记忆，这个过程能被视作是在新的信息片段到来时生成被存储的相关知识。



图三：记忆网络结构

3. **输出**，同时处理问题特征项两盒当前记忆并生成一个答案的特征向量。这里是推断必然发生的地方。在最简单的情况中，这部分能被用作一个针对所有被使用的记忆时段的排序函数，同时最高得分的支持记忆能够通过如下式子被检索：

$$o_1 = O_1(x, M) = argmax_i s_O(x, m_i)$$

其中s_O是一个对问题和一个记忆时段的内容之间的匹配度进行计分的函数。

4. **回复**，处理答案特征向量并生成一个被系统输出的自然语言的表述。理想情况下，一个输出一系列文本表征的 RNN 或者 LSTM 网络应该满足这个部分。

记忆神经网络中的核心创新在于将对记忆的读/写操作规范化成一个可区分的函数，因此允许了它本身经由剩下的神经网络实现的梯度下降而被训练。这在思想上和并行地发布的在神经图灵机上的工作（[3]中格雷福斯的工作）是相似的。记忆神经网络有效地将问答问题规划成两个步骤：找到与提问最相关的记忆片段，然后运用这些记忆去生成一个自然语言表述的答案，而这两个组件会在同一个损耗函数下一起被训练。正如我们将在第四章节中看到的那样啊，这种两步走的方法能够在文章文本在长度上增加时大幅改进 LSTMs 的性能。

4.2 弱监督的记忆神经网络

在训练数据中，[10]中所描述的记忆网络对一个提问需要对应的支持记忆。这就意味着记忆神经网络无法被应用到到其他广泛的问答任务中，因为每一个数据集

将需要被每个问题的支持情况所注释。[8]描述了另一个版本的记忆神经网络，它不需要这些信息。这个框架就像[8]中所描述的那样，我们鼓励读者去阅读原文获取更多细节。我们在下面展示了一个简要的概览。

图 4 展示了涉及到的网络的一层，这里有三个部分：

1. **输入端**：使得输入句子变成

$$x_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$$

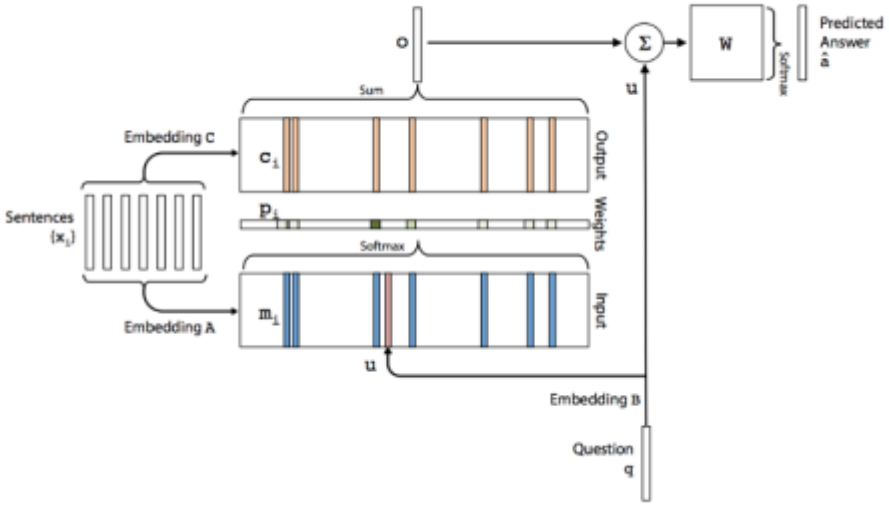


图 4：弱监督的记忆网络

使用用矩阵 A 嵌套句子中的每一个词汇。对于每一个句子，我们通过对句中所有词汇的嵌套进行求和计算出一个相应的记忆m_i。

$$m_i = \sum_j A x_{ij}$$

使用另一个矩阵 B 嵌套问题，并且我们计算问题和每一个记忆间的匹配度如下

$$p_i = softmax(u^T m_i) = softmax\left(q^T B^T \sum_j A x_{ij}\right)$$

2. **输出端**：每一个记忆向量有一个相应的输出向量c_i，用另一个嵌套矩阵 C 计算

$$c_i = \sum_j C x_{ij}$$

从记忆中产生的输出向量o是一个对c_i的求和，由输入得到的可能性向量进行赋权

$$o = \sum_i p_i c_i$$

3. **答案预测：**

$$a = softmax\big(W(o+u)\big)$$

像这样的一层相当于一个记忆查表。我们能够把很多这样的层叠加并利用$u^{k+1} = u^k + o^k$，最后使用矩阵W去预测一个答案。

[3]中的模型需要训练数据被支持记忆注释。大多数数据集没有这些信息。我们将弱监督的记忆网络运用到维基问答和 MC 测试数据集中。

接下来，我们将描述我们对这两个真实问答数据的记忆神经网络所做的扩展。

4.3 对维基问答任务的扩展

[10]和[8]中描述的记忆网络在具有小词汇量的数据集上训练和表现得很好。但是，在 bAbI 任务中当词汇量的规模增加到超过四十个词的时候，模型对每一次迭代要花费很长的时间。我们描述两种我们在维基问答数据集（[9]中的数据集）上用于训练弱监督的记忆网络的方法。这两种方法使用了故事语句的一些语义去剪除对应每个问题的情况的集合。剪枝（预处理的一种形式）在这两种方法中的确起了作用。在训练模型时很容易看到这种优势。把一个记忆网络应用到诸如维基问答这样的真实数据集上会花费较长的时间。第二，它去除了一些噪声情况，这将帮助到模型更有效地找到相关记忆。

1. **基于 WordVec 的：**我们使用预训练后的词向量。对于每一对（问题，情况），我们纳入每一个标志的词向量并计算问题和情况之间的余弦相似性。我现在有每一种情况的一个分数，并能够剪除低于某个确定下限的句子或者获取最高的 N 个情况。在一个每个问题有 100 种情况的数据集中，这在模型训练速度上带来了巨大的提高。

2. **基于部分语句（POS）的：**对于问题中的每个标志和故事语句中的每个标志，我们找到它们的 POS 标签。我们剪除那些不包括任何与问题共有的名词的情况。这是一种策略。基于数据集，我们能够选择一种更相关的策略并让模型训练得更快。

4.4 对 MC 测试任务的扩展

[8]中的弱监督记忆网络是被设计用于输出一个覆盖词汇量的可能性分配。对于 MC 测试数据集，我们对每个问题给出四个选项。一个挑战是能够用多个词汇处理回答。另一个挑战（改进）是用四个给定选项去做出更好的预测。

将选项纳入训练过程，我们用一个这种形式的计分函数：

$$p_i = u_k^T U^T U a_i$$

其中

$$u_k = o_{k-1} + u_{k-1}$$

a_i是第 i 个选项。我们将有最高分数的选项作为预测的回答。在训练中，我们用交叉熵损耗函数[2]。

这种计分函数的优势是各个选项现在也是训练的一部分，并且我们能在反向传播其错误。对于一个这种形式的一个选项：

$$x_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$$

然后我们计算它的嵌套如下：

$$\mathrm{a}_i = \sum_j D x_{ij}$$

其中 D 是选项相应的嵌套矩阵。

这允许我们同样可以对多个词汇的选项进行计分和排序。

5. 实验

5.1 bAbI 问答任务

作为第一项实验，我们训练了一个 LSTM 网络，它能阅读文章文本和一个问题并输出一个词汇作为回答。这个 LSTM 网络当它被喂给伴随着问题的一整篇文章时表现得很差[3]，但是，当它只是被喂给与问题相关的一些句子时它表现得非常好，在任务问答 1、问答 2、问答 3（详见表格 1）的测试集上获得了 100%的正确率。这表明了就像在记忆神经网络中所做的那样运用一个记忆组件到问答任务中的重要性。

我们使用了一个嵌套尺寸为 100 个词向量和 100 个输出的 LSTM 单元，它连接了随后是 Softmax 激活函数的 Dense 层。我们训练了网络，用 RMSProp 去最小化交叉熵损耗并从回答词汇到 LSTM 的最后输出的期间做反向传递。我们训练了这个网络 30 次迭代。

我们应用了[10]中所描述的受监督记忆神经网络，用词袋表示了输入情况和问题文本。我们也在情况被写入记忆中时应用了写时建模（详见[10]中 3.4 节）去记录相关顺序。这个网络能够学习在从记忆中捡取相关情况是去运用信息。我们固定了嵌套尺寸为 100，学习速率为 0.01 和幅度为 0.1，并且每次训练的迭代为 10 次。这个测试的正确率在表格 1 中给出。

表格 1：bAbI 问答任务的测试正确率

bAbI QA Task	LSTM w/ rel. stmts	LSTM w/ entire article	Supervised MemNN (k = 2)	Weakly sup. MemNN
QA1: Single Supporting Fact	100%	31.2%	100%	66.7%
QA2: Two Supporting Facts	100%	35.6%	77.5%	60.3%
QA3: Three Supporting Facts	100%	27.1%	42.9%	50.7%

我们也应用了[8]中所描述的弱监督记忆神经网络，用了三层和绑定每一层的权重到相同的输入/输出矩阵对上。我们纳入了论文中描述的位置编码矩阵。我们设置了所有论文中描述的参数值。值得注意是，受监督记忆神经网络在问答 3 中表现得很差，由于我们只考虑了两个相关的记忆（k=2）。相反地，弱监督记忆神经网络在这个特定的任务中表现得比受监督记忆神经网络更好，因为我们在弱网络中使用了 3 个输入/输出嵌入层，所以它能狗学习去识别多到 3 个支持记忆。

表格 2：在维基问答数据集上的性能

Wiki QA			
	Train Accuracy	Test Accuracy	Speedup
LSTM (w/ entire article)	84.2%	30.1%	-
WMemNN (Basic)	88.9%	44.4%	1x
WMemNN + POS Pruning	73.6%	46.5%	4x
WMemNN + WV Pruning	72.1%	45.7%	5x

5.2 维基问答任务

在测试了我们记忆神经网络和弱监督记忆神经在 bAbI 任务上的应用之后，我们在更复杂的维基问答数据集上训练了模型去分析它们在真实数据上的性能。维基问答数据要大很多（每篇文章有 100 到 1000 种情况，40000 多词汇），因此训练要花费更长的时间。因为训练时间的限制，我们把我们的训练集和测试集限制到每个问题的答案只有一个单词（其中大多数都是是/否类型的答案），丢弃了剩下的。另外，我们尝试着去降低训练时间同时提高正确率，通过使用 4.3 节中所描述的基于 POS 和词向量的剪枝技术去剪除输入文章中不相关的情况。结果展示在

表格 2 中。我们比较了弱监督记忆神经网络实现和基础的 LSTM 实现，它们都被喂给伴随着问题的整个维基百科文章，然后一个答案被从中抽取出来。LSTM 对少量的训练集产生了过拟合但是一点都不能泛化到文本样例中。一个普通 LSTM 没法很好地适应这项任务，因为一个完整的维基百科文章太复杂了，而且对 LSTM 来说在学习如何匹配问题和答案时有太多噪声了。弱监督记忆神经网络表现得比基础的 LSTM 更好。剪枝在训练时间上产生显著的提速，而且在测试性能上没有多少改变。

5.3 MC 测试任务

最后，我们应用了我们的弱监督记忆神经网络实现到 MC 测试数据集上。原文[5]描述了使用一种滑动窗口计分和一种基于距离计分相结合的基础方法我们重新实现了基础方法并能够重新产生文中所描述的正确率数值。接下来，我们训练了弱监督记忆神经网络模型在这个任务上，因为这个任务有多项选择答案而不是开放式答案组成，记忆神经网络模型需要被额外加强。值得注意的是，记忆神经网络（弱监督记忆神经网络）的输出是一个提取词的序列，这个序列被作为了答案。

我们尝试了两种处理多选项问题的方法：（i）第一，我们限制了我们的训练/测试集在答案只有一个单词的问题的范围。弱监督记忆神经网络模型（或者基础 LSTM）输出了一个覆盖整个词汇库的 Softmax 可能性向量。在四个答案中有最好可能性的答案被选为问题的答案。结果展示在表格 3 中。我们能够观察到基础的 LSTM 表现得比弱监督记忆神经网络更好。（ii）在我们第二种方法中，我们弱 4.4 节描述的那样扩展了弱监督记忆神经网络，使它能输出四个答案的排序结果，因此我们选出得分最高的一个。不幸的是，我们的弱监督记忆神经网络实现不能打败简单的基础版本。我们能够看到被扩展后的弱监督记忆神经网络模型几乎对整个数据集都过拟合了。我们尝试了基础的 L2 正则化，但是并不起作用。尝试可供替代的正则化方案应该改善这个问题，但是我们没有时间去找出方案来（详见 6.1 节：未来的工作）我们将在下一节中尝试错误分析。

表格 3：使用了 5.3 节所描述的两种方法的在 MC 测试数据集上的表现

	Approach 1: Restrict dataset		Approach 2: Incorporate choices into model			
Dataset	LSTM Test Accuracy	WMemNN (Basic) Test Accuracy	Baseline (SW + D)		WMemNN (Extended)	
			Train	Test	Train	Test
MC160	51.6%	45.2%	72.5%	66.2%	92.9%	36%
MC500	40.1%	36.5%	59.3%	56.7%	98.3%	34.2%

5.4 错误分析

我们的记忆神经网络和弱监督记忆神经网络实现表现优异 在最初的是哪个 bAbI 任务中，这能够判定针对大规模问答任务的记忆增强模型的性能，简单 LSTM 模型无法从一长篇文本中记忆所有事物去基于此正确地回答问题。但是弱监督记忆神经网络在真实数据集上的表现也不令人满意。在维基问答任务中，弱监督记忆神经网络模型表现得比基础的 LSTM 更好，但是性能依然比它在 bAbI 任务中要差。在 MC 测试任务中，弱监督记忆网络表现得比基础版更差。在人工检查了错误情况后，我们发现了这些错误的普遍原因：

1.**前指代解析**：错误最普遍的来源是在情况和回答包含了指代之前语句中提及的名词的代词的时候。在将代词输入弱监督记忆神经网络之前，解析文本中这样的代词应该带来优化。

2.**辨别同义词和上位词**：如果问题解释了输入文本的一部分或者使用了文本中出现过的词汇的同义词/上位词，弱监督记忆神经网络模型就没法处理这些，而且简单的基于名词/命名实体的剪枝在这种情况下适得其反。例如，对于文本“John plays baseball”，问题可能会是“Which sport does John play？”一种解决方法是使用预训练的词向量（比如[4]中的 word2vec 模型）去辨别文本和问题中的同义词和上位词并且表现出更加智能的剪枝。

3.**对复杂句子结构进行句法分析**：另一个重要问题发生在分析复杂句子的时候。例如，对于文本“Steve also like bananas, oranges and apples, but fish was his favorite.”和问题“What was Steve’s favorite food?”,模型选择了“apples”而不是“fish”。对于在弱监督记忆神经网络最后的输出层中的 W 矩阵而言，学习对比性的连接词在形如“X but Y”的句子中的作用几乎是不可能的。也许，如在[7]中所做的那样加入一个递归张量神经网络组件（RNTN）能够让模型更好地处理这样的复杂句子结构。

我们在附录 A 中提供了对 MC 测试任务中失败情况的例子的具体的错误分析。

6. 结论

我们研究了基于深度记忆神经网络的问答系统在仿真数据集和两个真实数据集上的表现。即使针对大规模问答任务使用记忆网络很明显是基于他们在仿真数据集上的表现的，但是让记忆网络在更加复杂的数据集上运作需要基本任务特定的特征提取和模型扩展。总之，首先使用传统的 NLP 特征提取有效地剪除输入、然后将修剪后的情况和问题通过一个记忆网络去获取答案的一个两步走的混合方法，似乎是在复杂问答任务中运作的最好的。

6.1 未来的工作

因为时间有限，我们在这项工作中剩下了少许研究方向没有探索。比如，在维基问答任务中，我们能在维基百科文本和问题中标记命名实体，并且只保留文章中那些包括出现在问题中的命名实体或者与它们很靠近的情况。这应该还一个提升训练速度和预测速度的廉价方法，同样也潜在地提升正确率，因为我们发现这个任务重的大多数问题只关注一个或者两个文本中的命名实体。

[8]中的在弱监督记忆神经网络上的工作没有触及生成多词答案的方面。对于任何一个真实的问答，或者对于模型去进行一场对话，系统必须生成有意义的句子。一种实现方法是加入一个在一个语言模型上预训练的 LSTM 到弱监督记忆神经网络的最后输出层，让其在回答中输出实用的句子。

相似地，对于 MC 测试任务，我们观察到简单的启发式的基础版表现得远比扩展后的弱监督记忆神经网络实现更好。将基础版纳入弱监督记忆神经网络应该会提升正确率。其次，弱监督记忆神经网络模型对这项任务重的训练数据集严重地过拟合，因此研究对模型的损耗函数进行正则化的作用是很有意义的。

参考文献

[1] Burges, Christopher JC. ”Towards the Machine Comprehension of Text: An Essay”. Microsoft Research Technical Report MSR-TR-2013-125, 2013, pdf, 2013.

[2] Fader, Anthony, Luke S. Zettlemoyer, and Oren Etzioni. ”Paraphrase-Driven Learning for Open Question Answering.” ACL (1). 2013.

[3] Graves, Alex, Greg Wayne, and Ivo Danihelka. ”Neural Turing Machines.” arXiv preprint arXiv:1410.5401 (2014).

[4] Mikolov, Tomas, et al. ”Distributed representations of words and phrases and their compositionality.” Advances in neural information processing systems. 2013.

[5] Richardson, Matthew, et al. ”MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text” EMNLP (2013)

[6] Smith, Noah A., et al. ”Question Generation as a Competitive Undergraduate Course Project” In Proceed- ings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge, Arlington, VA, 2008.

[7] Socher, Richard, et al. ”Recursive deep models for semantic compositionality over a sentiment treebank.” Proceedings of the conference on empirical methods in natural language processing (EMNLP). Vol. 1631. 2013.

[8] Sukhbaatar, Sainbayar, et al. ”Weakly Supervised Memory Networks” arXiv preprint arXiv:1503.08895 (2015).

[9] Weston, Jason, et al. ”Towards AI-complete question answering: A set of prerequisite toy tasks.” arXiv preprint arXiv:1502.05698 (2015).

[10] Weston, Jason, Sumit Chopra, and Antoine Bordes. ”Memory networks.” arXiv preprint arXiv:1410.3916 (2014).

附录 A MC 测试的错误分析

我们提供了少许弱监督记忆神经网络在 MC160 问答任务上的错误情况。对于每一个样例，我们只提供了输入文本中与问题有关的部分。每个问题有四个选项。正确答案用绿色标出，模型猜测的答案用红色标出。对这类错误的评论也在这里给出了。

Story: mc160.train.31

Ryan and Adam love to play basketball. They like it better than soccer and baseball. Their other friend, Jared, has his own basketball hoop. He got it for his ninth birthday. Ryan got a football for his birthday and Adam got a skateboard. They like their presents, but think the basketball hoop is better. They play basketball at Jared’s house with him and any other kids who show up. Alex and Brady come almost every day and Josh, Ty, and Max come sometimes. Next year, they all get to play on a basketball team. They get to play at their school. They are very excited about that and can’t wait to play on a real team. For now, they are practicing a lot and are trying to get really good. They play every day they can. They are trying to be as good as the NBA players they watch on TV. They dream of someday playing in the NBA. They are sure it is going to happen.

4: multiple: What did Jared get for his ninth birthday? A) a basketball hoop B) a baseball bat C) a football D) a skateboard

图 A.1：这里，答案是分成两句话的，并且第二句话包含了指代 Jared 的代词，因此模型没法将其提取出来。但是，它将表达“Ryan got a football for his birthday”匹配到了答案的表达“get for his ninth birthday”并且随后选择了“a football”作为答案。

Story: mc160.train.35

Once upon a time, there was a boy named Freddy. And Freddy loved his mom very much, and his mom loved him very much too. One day, Freddy went outside to ride his bike. On the way out, his mother told him, ”Remember to wear your helmet,” and Freddy grabbed his helmet and met his friends outside. When he was putting on his helmet, his friends told him, ”Helmets are for girls! You’re not cool if you wear a helmet!” Freddy thought about what his mom told him, but he wanted to be cool like his friends, and he took off his helmet.

1: one: Who does Freddy love? A) His friends B) His bicycle C) His mom D) His dad

图A.2：这个显然是一个简单的问题，但是模型做错了。所有三个关键词“Freddy, mom and love”都出现在输入的一个简单句中，但是弱监督记忆神经网络模型没法识别这个作为正确答案。

Story: mc160.train.36

There was a big race in town. Stephanie and Sarah were friends. Stephanie was faster than Sarah. On the day of the race, they wished each other good luck. Sarah tripped on a rock during the race. She cried but another one of her friends, Matt, helped her stand up. Stephanie cheered for her to finish after she crossed the line.

1: multiple: Based on the story, who likely won the race? A) Jane B) Matt C) Sarah D) Stephanie

图A.3：模型想要做对这个题目太难了，因为问题“who won the race？”是“after [Stephanie] crossed the line”这句话的引申义。模型在同一个句子中看到“Sarah”和“race”，因此输出了Sarah作为答案。

Story: mc160.train.43

Bailey and her friend Kara were bored one Saturday. It was a hot summer day. They didn’t want to stay inside any longer but they didn’t know what to do. They were tired of watching TV inside. Suddenly, Kara had an idea. She said, ”Bailey, we could make some money.” ”How?,” asked Bailey. ”Well, it is hot outside,” said Kara. ”People are thirsty out there. We could make money by making some lemonade and iced tea and have people pay for it.” ”That is a great idea,” answered Bailey, ”let’s do it!” Kara had made some iced tea with her mom earlier that day. She asked her mom permission to use it. Her mom said yes. She and Kara made two pitchers of lemonade. They got a cooler full of ice and made a sign so people knew what was for sale. Kara’s mom helped them get a table and chairs and set up out on the corner in their neighborhood. It was so hot out that people who saw their stand came to buy drinks right away. Their first visitors to their stand were their friends, Abby and Molly. In a half hour, they had to close their stand. They were all out of lemonade and iced tea. They had made a lot of money. They split the money and each got ten dollars. It was a great day.

3: multiple: Who help them set up their stand?

A) Abby’s mom B) Bailey’s mom C) Molly’s mom D) Kara’s mom

图A.4：在这次测试中，答案又包含了包含答案的输入句子的引申义。模型不能辨别“set up their stand”是“helped them get a table and chairs...”的引申义。它看到“Abby”和“stand”在同一个情况下并因此给“Abby’s mom”最高分。

Story: mc160.train.52

Sarah is a girl. Sarah has one brother. Sarah’s brother’s name is Timothy. Sarah has one sister. Sarah’s sister’s name is Annabelle. Their last name is MacGregor. One day Sarah went to the park with her brother Timothy. They swung on the swings for a short time. Then Annabelle came out and swung with them. They all sang some nice songs together. They all became very happy. Then Timothy’s friend came. Timothy liked his friend very much. Timothy went off the swing and went away with his friend. Then Annabelle and Sarah felt very very sad. Happily then Annabelle and Sarah’s friend came. Their friend’s name was Kate Smith. She was the same age as Sarah. They wanted to go to the slide together. So they went to the slide and played for a long time. Then Annabelle became happy. And Sarah also became happy. Then they went home together and had some food.

2: multiple: How many brothers and sisters does Sarah have? A) 1 B) 2 C) 0 D) 3

图A.5：这个问题对于弱监督记忆神经网络是极其困难的，因为它必须掌握数字的加法，当“Sarah has one brother”和“Sarah has one sister”出现在输入文本中的不同地方。

Story: mc160.train.65

Grace wants to play Frisbee. She goes to her store to buy a Frisbee. She picks out a red Frisbee. It is small enough to fit in her hand. It costs 75 cents. She buys it. She leaves the store. When Grace gets home, she has no one to play with. She looks for her friend Susan. Susan is not at home. She looks for her friend Jeff. Jeff is not allowed to go outside. Grace finds a dog named Ginger. Ginger loves to play frisbee. Grace tosses the frisbee to Ginger. Ginger catches it in her mouth. Ginger brings the frisbee back to Grace. Grace tosses the frisbee again. Ginger jumps up in the air and catches it. Grace throws the Frisbee one more time. The Frisbee lands in a tree. Grace is too short to reach the Frisbee. Grace pets Ginger and tells her that she is a good girl. Grace takes Ginger home. They eat cookies. The next day, they come back to the park. They get their Frisbee back. They play again.

2: multiple: Who did Grace look for when she left the store? A) Susan and Jeff B) Jeff and Grace C) Susan and Ginger D) Ginger and Jeff

图A.6：弱监督记忆神经网络又是很难答对这个题的，因为他不能分辨Ginger是一条狗而Susan和Jeff是Grace的朋友。