神经网络序列到序列的学习

March 12, 2018

Abstract

深度神经网络（DNNs）是强大的模型，在困难的学习任务上取得了出色的表现。尽管DNNs在很大的标签训练集上的方法，对序列结构做出最小假设。我们的方法使用一个多层长短期记忆网络（LSTM）来把输入序列映射到一个固定维度的向量，然后另一个深度LSTM网络从这个向量中解码出目标序列。我们的主要结果是基于WMT-14数据集做的英语到法语的翻译，这个由LSTM产生的结果在整个测试数据集上达到了34.8的BLEU得分，BLEU得分是越M处理长句子并不困难。作为比较，一个基于短语的SMT系统在相同的数据集上获得了33.3的BLEU得分。当我们使用LSTM来重跑由之 前SMT系统产生的1000个假设，BLEU得分增加到了36.5，很接近先前的最新状态。LST-M也能学习合理 的短语和句子表述，对词序很敏感，对主动和被动语态相对不变（不太敏感）。我们发现把源句子（-> d,c,b,a）输入，LSTM的表现明显提高了，因 为这样做引入了源和目标语句之间的许多短期依赖关系，从而使优化

1 介绍

深度神经网络（DNNs）是非常强大的机器学习模型，注意在诸如语音识别[13,7]和视觉对象识别[19,6,21,20]等难题上实现卓越的性能。Dnns功能强大，因为对于少量步骤，它们可以执行任意并行计算。Dn够的信息来指定网络的参数，就可以用监督反向传播来训练。因此，如果存在获得良好结果的大数据集的参数

尽管它们具有灵活性和强大的功能，但DNNs只能用于输入和目标可以用维数固定的矢量进行合理编码的问题和机器翻译是序列问题。同样，问题回答也可以被看作是将表示问题的单词序列映射到表示答案的单词序列。

序列对DNNs提出了挑战，因为它们要求输入和输出的维度是已知的并且是固定的。在本文中，我们展示了一个简单长短期记忆（LSTM）体系结构[16]的直接应用可以解决序列问题的一般序列问题。这个想法是STM来读取输入序列，一次一个时间步，以获得大的固定维向量表示，然后使用另一个从该向量提取输出序列STM本质上是一个递归的神经网络语言模型[28,23,30]，除了它是以输入序列为条件的。由于输入和它们相应输M成功处理 长时间依赖关系数据的能力使其成为该应用的自然选择。

已经有许多相关的尝试来解决用神经网络对序列学习问题进行排序的一般序列。我们的方法与Kalch-brenner和Blunsom密切相关[18]，他们首先将整个输入句子映射为向量，并且与-Cho等人非常相似[5]。Graves[10]引入了一种新颖的可区分关注机制，它允许神经网络将注意力集中在输入的部分，并且这一想法的一个优雅变体被Bahdanau等人成功应用于机器翻译[2]。连接主义序列分类是将序列映射到具有神经网络的序列的另一种流行技术，尽管它假设输入和输出之间是单调对齐的[11]。

这项工作的主要成果如下。在WMT’14英语到法语翻译任务中，我们使用简单的从左到右的波束搜索解码器，通过从5个深层（全部380m参数）的集合直接提取翻译来获得34.81的BLEU分数。这是迄今为止通过大33.30[29]。这个34.81BLEU得分是由一个词汇为80k词的词条实现的，因此只要参考翻译包含一个未被这80k涵

最后，我们使用LSTM在相同的任务中重新公开了SMT基线的1000个最佳列表[29]。通过这样做，我们获得了

令人惊讶的是，尽管最近有其他研究人员使用相关体系结构[26]，但LSTM并没有遭受很长的句子。我们能够

LSTM的一个有用属性是它学习将可变长度的输入句子映射为固定维向量表示。由于翻译往往是对源句的解释STM找到能够捕捉其含义的句子表示法，因为具有相似含义的句子彼此接近，而不同的句子含义将会很远。定性

2 模型

递归神经网络（RNN）[31,28]是前馈神经网络对序列的自然泛化。给定输入序列（x1，...，xt），标准RNN通过迭代以下等式来计算输出序列（y1，...，yt）：

$$h_t = sigm\left(W^{hx}x_t + W^{hh}h_{t-1}\right)$$

$$y_t = W^{yh}h_t$$

只要提前知道输出之间的对齐，RNN就可以轻松地将序列映射到序列。然而，对于复杂和非单调关系的输入和输出序列长度不同的问题，还不清楚如何解决。

一个简单的通用序列学习策略是将输入序列映射到一个固定大小的载体上，然后用另一个RNN将载体映射到目标序列（Cho等人[5]也采用了这种方法）。尽管原则上它可以工作，因为RNN提供了所有的相关信息，但由于长期的相关性[14，4]（图1）[16,15]，难以训练RNNs。然而，长时间的短期M可能会成功。

LSTM的目标是估计条件概率p（y1，...，yt’｜x1，...，xt），其中（x1，...，xt）是输入序列，并且y. . . ，yT′是它的相应输出序列，其长度可能不同于T。LSTM通过首先获得由LSTM的最后一个隐藏状态给出的输入序列（x1，...，xt）的固定维数表示，然后计算y1，. . . ，yT′的概率，来计算该条件概率。y1，. . . ，yT′具有标准的LSTM-lm公式，其初始隐藏状态被设置为表示x1，...，xt

$$p\left(y_1,...,y_{T'}|x_1,...,x_T\right) = \prod_{t=1}^{T'} p\left(y_t|v,y_1,...,y_{t-1}\right)$$

在该等式中，词汇表中的每个词都用softmax表示每个p（yt｜v，y1，...，yt-1）分布。我们使用Graves [10]中的LSTM公式。请注意，我们需要每个句子以特殊的句尾符号“<EOS>”结尾，这使模型STM计算“a”，“b”，“c”，“<EOS>”的表示，然后用这个表示来计算“w”，“x”“y”，“z”，“<EOS>”。

我们的实际模型在三个重要方面与上述描述不同。首先，我们使用了两个不同的lstms：一个用于输入序列，另一个用于输出序列，因为这样可以以可忽略的计算成本增加数字模型参数，并使stm [18]。 其次，我们发现深层表现明显优于浅层，所以我们选择了四层结构。
第三，我们发现它对于改变输入句子的单词顺序非常有价值。例如，将句子a，b，c映射到句子α，β，γ，LM要求映射c，b，a到α，β，γ，其中α，β，γ是a，b，c的平移。这样，α接近于α，β接近于β，等等，gd可以很容易地在输入和输出之间“建立通信”。 我们发现这个简单的数据转换可以大大提升lstm的性能。

3 实验

我们用两种方法将我们的模型应用到WMT’14英语翻译为法语的MT任务。 我们使用它直接翻译输入句子而不使用我们报告这些翻译方法的准确性，呈现样本翻译，并将所得到的句子表示形象化。

3.1 数据集详细信息

我们用wmt’14英语到法语数据集。 我们对包含348M法语单词和304M英文单词的12M语音子集训练了我们的模型我们选择了这个翻译任务和这个特定的训练集子集，因为公开提供了一个标记化的训练和测试集以及来自baselineSMT [29]的1000个最佳列表。

因为典型的神经语言模型依赖于每个单词的矢量表示，所以我们为这两种语言使用了固定的词汇表。
我们使用源语言中最频繁的单词为160,000个，目标语言中最常用的单词为80,000个。
每一个词外的单词都被一个特殊的“UNK”标记取代。

3.2 解码和重新编码

我们的实验的核心涉及在许多句子对上训练大量的深层文体。 我们通过一个给定源语S的正确翻译T的最大对数

$$1/|s| \sum_{(T,S)\epsilon s} \log p(T|S)$$

s是训练集。一旦培训完成，我们根据lstm找到最可能的翻译来制作翻译：

$$\hat{T} = \arg\max_{T} p(T|S)$$

我们使用简单的从左到右的波束搜索解码器来搜索最可能的翻译，该解码器保持少量的部分假设，其中部分
在每个时间步，我们用词汇表中的每个可能的词来扩展束中的每个部分假设。 这大大增加了假设的数量，所以
只要“<eos>”符号附加到假设上，它就会从波束中移除并被添加到完整假设集合中。
而这个解码器是近似的，它的实现很简单。 有趣的是，即使光束大小为1，我们的系统也能很好地工作，而光
我们还使用LSTM来重新调整由基线系统生成的1000个最佳列表[29]。 Torescore是一个n-
best列表，我们用我们的LSTM计算了每个假设的对数概率，并用他们的得分和L-
STM的得分取平均值。

3.3 颠倒源句子

虽然LSTM能够解决长期依赖的问题，但我们发现当源句子颠倒时（目标句子未被反转），LST-
M学习得更好。 通过这样做，LSTM的测试困惑度从5.8下降到4.7，并且其解码翻译的测试BLEU成绩从25.9上升
虽然我们对这种现象没有完整的解释，但我们认为这是由于对数据集引入了许多短期依赖性。
通常情况下，当我们将源语句与目标语句连接起来时，源语句中的每个单词与目标语句中的相应单词很远。
因此，这个问题有很大的“最小时滞”[17]。 通过反转源句子中的单词，源语言和目标语言中相应单词之间的
然而，源语言中的前几个单词现在与目标语言中的前几个单词非常接近，因此该问题的最小时滞大大减少了。
因此，反向传播在源语句与目标语句之间“建立沟通”的时间较为容易，从而导致整体表现的显着改善。
最初，我们认为，逆转输入句子只会导致对目标句子早期部分的更有信心的预测，并且会导致对后面部分的
s对反向源语句的训练比对原语句子训练的LSTM的长句要好得多（见3.7节），这表明在L-
STMs中反转输入句子的结果具有更好的内存利用率。

3.4 训练详细信息

我们发现LSTM模型相当容易培训。 我们使用了深度为4层的LSTM，每层1000个单元格和1000维的单词嵌入，输
我们发现深层的LSTM显着优于LSTM，其中每个附加层减少了近10％的困惑，可能是由于它们的隐藏状态更大。
我们在每个输出中使用了超过80,000字的朴素软尺。 所得到的LTM具有380M参数，其中64M是纯循环连接（32M

- 我们用-0.08和0.08之间的均匀分布初始化了所有的参数
- 我们使用没有动量的随机梯度下降，固定学习率为0.7。经过5个epoch后，我们没半个e-
  poch就减少一半学习率。 我们训练了总共7.5个epoch的模型

- 我们使用批次128个序列作为梯度，并分割批次的大小（即128）。
- 虽然LSTM倾向于不受渐变梯度问题的困扰，但它们可以有渐变梯度。 因此，我们通过在其范数超出阈值时
  $||g||_2$，g是梯度除以128。如果$s > 5$，我们设置$g = \frac{5g}{s}$。
- 不同的句子有不同的长度。 大多数句子短（例如，长度20-30），但是一些句子长（例如，长度> 100），因此随机选择128个句子的小批量将具有许多短句子和很少长句子，结果，大部分句子mini-batch中的计算被浪费了。 为了解决这个问题，我们提出了一个minibatch中的所有句子大致相同的长度

3.5 并行

深度LSTM的C++实现使用单个GPU上一节中的配置，每秒处理大约1,700字的速度。这对于我们的目的来说太慢了
GPU机器来并行化我们的模型。 LSTM的每一层都在不同的GPU上执行，并在计算后立即将其激活传达给下一个GP
我们的模型有4层LSTM，每层LSTM都位于单独的GPU上。 剩余的4个GPU用于并行化-softmax，因此每个GPU负责乘以1000×20000矩阵。 由此产生的实现速度达到每秒6,300（包括英文和法文）的

3.6 实验结果

我们使用BLEU评分[24]来评估翻译质量。 我们使用multi-bleu.pl1对经过校正的预测和基础事实进行了计算。
这种评估BELU得分的方法与[5]和[2]一致，并且再现了[33]的33.3得分。然而，如果我们评估[9]的艺术体系状
downloadedfromstatmt.org\matrix）以这种方式，我们得到37.0，这大于报告的35.8reportedbystatmt.org\matrix。

结果列在表1和表2中。我们的最佳结果是通过随机初始化和小型随机顺序不同的LSTM集合获得的。 尽管LSTM集合的解码翻译并没有打破现有技术水平，但这是纯粹的神经翻译系统首次在大型M
通过重新评估基线系统的1000个最佳列表，LSTM在先前技术状态的0.5BLEU点内。

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	34.81

表1：LSTM在WMT’14英语到法语测试集（ntst14）上的表现。 请注意，具有2号光束的5个LSTM的集合比具有12个光束的单个LSTM便宜。

Method	test BLEU score (ntst
Baseline System [29]	33.30
Cho et al. [5]	34.54
State of the art [9]	37.0
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	36.5
Oracle Rescoring of the Baseline 1000-best lists	~45

表2：在WMT’14 Englishto French测试集（ntst14）上使用神经网络和SMT系统的方法

3.7 在长句子上的表现

我们惊讶地发现LSTM在长句中表现良好，这在图3中定量地显示出来。表3给出了几个长句及其翻译的例子。

3.8 模型分析

我们模型的一个有吸引力的特征是它能够将一系列词汇转化为一个固定的维度向量。
图2显示了一些学习的表示。 该图清楚地表明，表示对词的顺序敏感，而对用主动语音替换主动语音相当不敏
二维投影使用PCA获得。

4 相关工作

神经网络应用于机器翻译有大量的工作。 到目前为止，在MT任务中应用RNN-语言模型（RNNLM）[23]或前馈神
MT基线[22]，可靠地提高翻译质量。

最近，研究人员开始研究如何将源语言信息纳入NNLM。 这项工作的例子包括Auli等人。[1]，他们将NNLM与
Devlin等人 [8]采用了类似的方法，但他们将NNLM并入MT系统的解码器中，并使用解码器的对齐信息为NNLM提
他们的方法非常成功，并且在基线之后取得了很大的改进。

我们的工作与Kalchbrenner和Blunsom密切相关[18]，他们首先将输入语句映射到一个向量中，然后再回到
[5]使用类似LSTM的RNN体系结构将语句映射到向量中并返回，尽管它们的主要焦点是将其神经网络集成到SMT系
Bahdanau等人。 [2]也试图用神经网络直接翻译，这种神经网络使用注意机制来克服-
Cho等人所经历的长句的不良表现。 [5]并取得了令人鼓舞的成果。同样，Pouget-
Abadie et al。 [26]试图解决Cho等人的记忆问题。 [5]通过翻译产生平滑翻译的源句子片断，这类似于基于短

端到端培训也是Hermann等人关注的焦点。 [12]，其模型代表前馈网络的输入和输出，并将它们映射到类似
然而，他们的方法不能直接生成翻译：为了获得翻译，他们需要在预先计算的句子数据库中查找最接近的向量

5 结论

在这项工作中，我们展示了一个词汇量有限的深层LSTM可以超越标准的基于SMT的系统，该系统在大型MT任务中
我们简单的基于LSTM的MT方法的成功表明，它应该很好地解决许多其他序列学习问题，只要他们有足够的训练

我们对翻译源句中的单词所获得的改进程度感到惊讶。 我们的结论是，找到一个具有最大数量短期依赖性
特别是，尽管我们无法针对非逆向翻译问题（如图1所示）训练标准RNN，但我们认为，当源句子被逆转时，标
N应易于训练（尽管我们没有通过实验验证）。

我们也对LSTM正确翻译非常长的句子的能力感到惊讶。 我们最初确信，由于记忆力有限，LST-
M在长句中会失败，其他研究人员报告称长句的表现不佳，其模型与我们的类似[5,2,26]。
然而，对逆向数据集进行训练的LSTMs难以翻译冗长的句子。

最重要的是，我们证明了一个简单，直接和相对未优化的方法可以超越成熟的SMT系统，因此进一步的工作
这些结果表明，我们的方法可能会在其他具有挑战性的序列问题上做得很好。

6 感谢

We thank Samy Bengio, Jeff Dean, Matthieu Devin, Geoffrey Hinton, NalKalchbrenner, Thang Luong, Wolf-gang Macherey, Rajat Monga, Vincent Vanhoucke, Peng Xu, Wojciech Zaremba, and the Google Brain teamfor useful comments and discussions。

References

[1] M. Auli, M. Galley, C. Quirk, and G. Zweig. Joint language and translation modeling with recurrentneural networks. InEMNLP, 2013.