通过编码器-译码器方式的神经机器翻译

摘要

神经机器翻译是一种只依赖神经网络较新的统计机器翻译。神经机器翻译模型通常由一个编码器和译码器组成。编码器从可变长度的输入中抽取固定长度的内容来代表输入，译码器从这个表示中生成一个正确的翻译。在这篇论文中，作者分析了两个神经机器翻译模型的性质：RNN 编码器-译码器和一个新提出的门控循环卷积神经网络 GRU。作者发现，在没有未知单词的短句中神经机器翻译表现较好，但是当句子长度增加或未知单词增多时，它的表现迅速下降。而且，作者发现他们提出的 GRU 会自动学习句子的语法结构。

一、导论

神经机器翻译受到了最近深度表示学习趋势的启发，它使用的所有神经网络模型都由编码器和解码器组成。

神经机器翻译模型只需要传统统计机器翻译（SMT）模型所需内存的一小部分。作者为本文训练的模型总共只需要 500MB 的内存，这与现有的往往需要几十 GB 内存的 SMT 系统形成鲜明对比。这使得神经机器翻译在实践中颇具吸引力。此外，与传统的翻译系统不同，神经翻译模型的每一个组件都是联合训练的，以最大化翻译性能。

优点：

内存需求少，组成模型的组件同时训练

由于这种方法相对较新，在分析这些模型的性质和行为方面的工作还不是很多。例如：这种方法表现更好的句子的属性是什么?源/目标词汇的选择如何影响翻译效果?在何种情况下神经机器翻译失败?

理解这种新的神经机器翻译方法的性质和行为对于确定未来的研究方向至关重要。此外，了解神经机器翻译的弱点和优势可能会促进更好的方式来集成 SMT 和神经机器翻译系统。

本文分析了两种神经机器翻译模型。一种模型是 RNN 编码器-译码器就是其中之一。另一种模型将 RNN 编码器-译码器模型中的编码器替换为一种新型的神经网络——门控递归卷积神网络（grConv）。作者通过将法语翻译到英语的任务来对这两个模型进行评估。

分析表明，随着源句长度的增加，神经机器翻译模型的性能迅速下降。此外，作者还发现词汇量对翻译性能有很大影响。尽管如此，这两个模型在大部分时间内都能生成正确的译

文。此外，新提出的 grConv 模型能够在无监督情况下学习源语言上的一种语法结构。

无监督学习：只有特征没有标签，是一种机器学习的训练方式，本质上是一个统计手段，在没有标签的数据里可以发现潜在的一些结构的一种训练方式。可以用于用户喜好推荐、生成文字序列等。

二、处理变成序列的神经网络

含门控隐层神经元的循环神经网络

循环神经网络通过隐藏状态来处理可变长度序列。在每个时间步 t，隐藏状态 h (t)更新为 h (t) = f(h(t-1), xt)，f 为激活函数。f 通常只需对输入向量进行线性变换，然后求和，并应用一个 sigmoid 函数。

RNN 可以通过学习下一个输入上的分布来有效地处理变长序列输入的分布。例如，对于一个长度为 K 的向量，其分布可以通过输出为

$$p(x_{t,j} = 1 \mid \mathbf{x}_{t-1}, \dots, \mathbf{x}_1) = \frac{\exp\left(\mathbf{w}_j \mathbf{h}_{\langle t \rangle}\right)}{\sum_{j'=1}^{K} \exp\left(\mathbf{w}_{j'} \mathbf{h}_{\langle t \rangle}\right)}$$

的 RNN 来学习，其中 wj 为权重矩阵 W 的行。共同分布的结果如下

$$p(x) = \prod_{t=1}^{T} p(x_t \mid x_{t-1}, \dots, x_1).$$

最近提出了对于 RNNs 一种新的激活函数。新的激活函数增加了两个门控单元（reset，r 和 updata，z）。每个门依赖之前的隐藏状态 h(t-1)和当前输入 xt 来控制信息的流动。这让人想起 LSTM。对于本文的剩余部分将总是使用这种新的激活函数。

门控循环卷积神经网络

门控递归卷积神经网络的每层参数都是共享的。在本节中，作者引入了一个二进制卷积神经网络，它的权重递归地应用于输入序列，直到它输出一个固定长度的向量。除了通常的卷积结构之外，作者提出使用前面提到的门控机制，它允许递归网络动态地学习源语句的结构。

令 x=(x1, x2, ⋯, xT)为输入序列，其中 xt∈Rd。本文提出的门控递归卷积神经网络（grConv）由 4 个权重矩阵 Wl，Wr，Gl 和 Gr 组成。在每个循环层 t∈[1 , T-1]，第 j 个隐藏单元 h(t) j 的为 ωc ˜h (t) j + ωlh (t-1) j-1 + ωrh (jt-1)，其中 ωc，ωl 和 ωr 是一个门的值，且和为 1。. 隐藏单元初始化为 H (0) J = Uxj，其中，U 将输入投影到一个隐藏空间。

新的激活函数$\tilde{h}_j^{(t)}$的计算如下：

$$\tilde{h}_j^{(t)} = \phi\left(\mathbf{W}^l h_{j-1}^{(t)} + \mathbf{W}^r h_j^{(t)}\right)$$

式中：φ 为单元非线性。

门控系数 ω 的计算公式为：

$$\begin{bmatrix} \omega_c \\ \omega_l \\ \omega_r \end{bmatrix} = \frac{1}{Z}\exp\left(\mathbf{G}^l h_{j-1}^{(t)} + \mathbf{G}^r h_j^{(t)}\right)$$

且

$$Z = \sum_{k=1}^{3}\left[\exp\left(\mathbf{G}^l h_{j-1}^{(t)} + \mathbf{G}^r h_j^{(t)}\right)\right]_k$$

根据这种激活函数，递循环层上单个节点的激活可以在从左、右子节点计算的新激活、从左子节点计算的激活或从右子节点计算的激活之中进行选择。这种选择允许循环卷积的整体结构根据输入样本自适应地改变。

在这方面，我们甚至可以认为本文提出的 grConv 是一种无监督的解析。如果考虑门控单元做出硬判决的情况，即 ω 遵循 1 - of - K 编码，很容易看出网络适应输入，形成树状结构。然而，对该模型学习到的结构还需进行进一步研究。

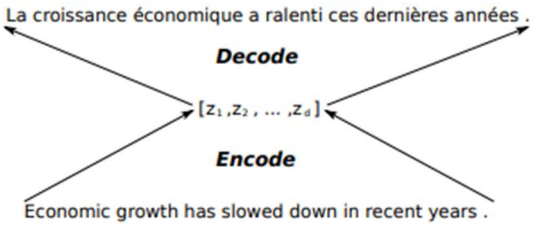
三、纯神经机器翻译

编码器-解码器方法

翻译任务可以从机器学习的角度理解为：给定源句 e，学习目标句（翻译）f 的条件分布 p(f |e)。一旦条件分布被一个模型学习到，就可以使用该模型对给定源句的目标句进行直接采 样，或者通过实际采样或者使用(近似)搜索算法找到分布的最大值。

许多论文提出使用神经网络直接从双语平行语料库中学习条件分布。例如，有人提出了一种使用卷积 n-gram 模型来提取源句子的固定长度向量的方法，该方法使用 RNN 增强的反卷积 n-gram 模型进行解码。还有人使用带有 LSTM 单元的 RNN 对源语句进行编码，并从最后一个隐藏状态开始，对目标语句进行解码。类似地，有人提出使用 RNN 对一对源和目标短语进行编码和解码。

所有这些工作的核心是编码器-解码器架构。编码器处理可变长度的输入（源语句）并构建固定长度的向量表示。解码器根据该向量表示生成一个可变长度的序列（目标句子）。



在之前，这种编码器-解码器方法主要作为现有统计机器翻译（STM）系统的一部分。该方法用于对 SMT 系统生成的 n – best list 进行重排序，还有人使用该方法为已有的短语表提供额外的评分。

在本文中，作者重点分析了在两种模型配置下的直接翻译性能。在这两个模型中，都使用一个门控隐藏单元，因为这是唯一不需要非平凡的方法来确定目标长度的选择之一。第一个模型将使用带有隐藏门控单元的 RNN 作为编码器，第二个模型使用门控循环卷积神经网络（grConv）。这样做的目的是通过 BLEU 来评估编码器-解码器方法在翻译表现上的归纳偏差。

四、实验设置

数据集

作者在英译法翻译任务上评估了编码器-解码器模型。使用的数据集是从 Europarl（61M）、news comment （5.5M）、UN（421M）以及两个分别为 90M 和 780M word 的爬虫语料库中选取的一组 348M 的双语平行语料库。神经机器翻译模型的性能在 news - test2012、news - test2013 和 news - test2014 (3000 行/组)上进行了测试。在与 SMT 系统进行对比时，使用 news - test2012 和 news - test2013 作为对 SMT 系统进行调优的开发集，news - test2014 作为测试集。

出于计算效率的考虑，我们只使用英语和法语句子长度不超过 30 个单词的句子对来训练神经网络。此外，我们只使用了英语和法语中最常用的 3 万个单词，所有其他的稀有词被认为是未知的，并映射到一个特殊的标记[UNK]。

模型

我们训练了两个模型：RNN Encoder-Decoder (RNNenc)和新提出的门控递归卷积神经网络(gr Conv)。这两个模型都使用一个带有门控隐单元的 RNN 作为解码器。

我们使用带有 AdaDelta 的小批量数据随机梯度下降来训练这两个模型。我们将平方权重矩阵(转移矩阵)初始化为正交矩阵，其谱半径在 RNNenc 中设置为 1，在 grConv 中设置为 0.4。RNNenc 中使用 tanh 激活函数，grConv 使用 relu 激活函数。

grConv 有 2000 个隐层神经元，RNNenc 有 1000 个隐层神经元。两种情况下词输入均为 620 维。两个模型均训练了约 110 小时，分别相当于 grConv 和 RNNenc 的 296144 次更新和 846322 次更新。

使用 Beam-Search 进行翻译

作者使用束搜索的一种基本形式来寻找是条件概率最大化的翻译。在解码器的每个时间步，我们保留对数概率最高的 s 个候选平移，其中 s = 10 为波束宽度。在波束 搜索过程中，我们排除了任何包含未知字的假 设。对于从最高评分候选者中选择的每个序列 结束符号，将波束宽度减小一个，直到波束宽 度达到零。 提出了 RNN 下波束搜索(近似)寻找最大对数概率序列的方法，并在(Graves , 2012)和(布 朗热- Lewandowski 等, 2013)中成功应用。最近，(Sutskever 等, 2014)的作者发现这种方法在基于 LSTM 单元的纯神经机器翻译中是有效的。

五、结果与分析

定量分析

在本文中，我们感兴趣的是神经机器翻译模型的性质。具体来说，翻译质量与源句/目标句的长度以及每个源句/目标句中模型未知词的数量有关。

反映翻译性能的 BLEU 随句子长度的变化情况如何评分。显然，这两个模型在短句上表现相对较好，但随着句子长度的增加而明显下降。

正如预期的那样，随着未知词数量的增加，性能迅速下降。这表明增加神经机器翻译系统输入的词汇量将是未来的一个重要挑战。

下表展示了使用两个模型和基于基线短语的 SMT 系统获得的翻译性能。显然基于短语的 SMT 系统仍然表现出优于所提出的纯神经机器翻译系统的性能，但是可以看到在特定的条件（源句和参考句中都没有未知词）下，这种差异明显减小。此外，如果只考虑短句（每句 10 ~ 20 个词），这种差异会进一步降低。

	Model	Development	Test
All	RNNenc	13.15	13.92
	grConv	9.97	9.97
	Moses	30.64	33.30
	Moses+RNNenc*	31.48	34.64
	Moses+LSTM°	32	35.65
No UNK	RNNenc	21.01	23.45
	grConv	17.19	18.22
	Moses	32.77	35.63

(a) All Lengths

	Model	Development	Test
All	RNNenc	19.12	20.99
	grConv	16.60	17.50
	Moses	28.92	32.00
No UNK	RNNenc	24.73	27.03
	grConv	21.74	22.94
	Moses	32.20	35.40

(b) 10–20 Words

此外，还可以将神经机器翻译模型与现有的基于短语的系统以提高整体翻译性能。这一分析表明，目前的神经翻译方法在处理长句方面存在不足。最明显的解释假设是固定长度的

向量表示没有足够的能力来编码结构和意义复杂的长句。为了编码一个可变长度的序列，神经网络可能会"牺牲"输入句子中的一些重要主题以记住其他主题。

这与传统的基于短语的机器翻译系统形成了鲜明的对比。在相同的数据集（为语言模型添加额外的单语数据）上训练的传统系统倾向于在更长的句子上获得更高的 BLEU 得分。

事实上，在句子和参考译文在 10~20 个单词之间，使用没有未知词的句子的情况下，RNNenc 和 Moses 在测试集上的 BLEU 得分分别为 27.81 和 33.08。值得注意的是，当使用多达 50 个单词的句子来训练这些模型时，观察到了类似的趋势。

定性分析

尽管 BLEU 评分被用作评价机器翻译系统性能的标准度量，但它并不是完美度量。因此，作者还展示了从 RNNenc 和 grConv 两个模型生成的一些实际翻译。

- 论文试图解决什么问题？

本论文的主要目的是对比 RNN Encoder-Decoder 和新提出的门控递归卷积神经网络

- 这是否是一个新的问题？

神经机器翻译方法相对较新，在分析这些模型的性质和行为方面的工作还不是很多

- 这篇文章要验证一个什么科学假设？

在循环神经网络中，通过使用门控机制可以有效地捕捉时间序列数据中的长期和短期依赖关系，同时解决梯度消失问题。

- 有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

门控机制的研究：GRU 的核心是门控机制，通过控制信息的流动来捕捉时间序列数据中的依赖关系。因此，与门控机制相关的研究是 GRU 相关研究的一个重要方向。

循环神经网络的应用研究：GRU 是循环神经网络中的一种单元，其应用在许多领域中都有研究。这些应用包括语音识别、自然语言处理、图像处理、时间序列预测等。因此，循环神经网络的应用研究也是 GRU 相关研究的一个重要方向。

长期依赖关系和短期依赖关系的研究：GRU 通过使用门控机制可以捕捉时间序列数据中的长期和短期依赖关系。因此，对于长期依赖关系和短期依赖关系的研究也是 GRU 相关研究的一个重要方向。

本论文作者：Yoshua Bengio

- 论文中提到的解决方案之关键是什么？

门控递归卷积结构

- 论文中的实验是如何设计的？

在定性分析和定量分析之间来进行对比

- 用于定量评估的数据集是什么？代码有没有开源？

使用的数据集是从 Europarl（61M）、news comment （5.5M）、UN（421M）以及两个分别为 90M 和 780M word 的爬虫语料库中选取的一组 348M 的双语平行语料库。神经机器翻译模型的性能在 news - test2012、news - test2013 和 news - test2014 (3000 行/组)上进行了测试。在与 SMT 系统进行对

比时，使用 news - test2012 和 news - test2013 作为对 SMT 系统进行调优的开发集，news - test2014 作为测试集；代码开源

- 论文中的实验及结果有没有很好地支持需要验证的科学假设？

在没有未知单词的短句中神经机器翻译表现较好，但是当句子长度增加或未知单词增多时，它的表现迅速下降。而且，作者发现他们提出的 GRU 会自动学习句子的语法结构。

- 这篇论文到底有什么贡献？

对比 RNN Encoder-Decoder 和门控递归卷积神经网络 GRU，从而掌握各自的适用条件

- 下一步呢？有什么工作可以继续深入？

神经机器翻译的性能受句子长度的影响很大，可以在这方面进行深入；GRU 可以用来探索语法结构；集成 SMT 和神经机器翻译系统