

基于深度可分离卷积的汉越神经机器翻译

徐 毓, 赖 华*, 余正涛, 高盛祥, 文永华

(昆明理工大学信息工程与自动化学院, 云南省人工智能重点实验室, 云南 昆明 650500)

摘要: 在汉越神经机器翻译中, 由于汉越平行语料稀少, 使得数据稀疏问题十分严重, 极大地影响了模型的翻译效果. 为了提升数据稀疏情况下的汉越神经机器翻译性能, 提出一种基于深度可分离卷积的汉越神经机器翻译方法. 该方法根据越南语的语言特点, 将越南语切分为词、音节、字符、子词 4 种不同的粒度并利用深度可分离卷积改进神经机器翻译模型, 通过增加深度可分离卷积神经网络, 对模型输入的不同粒度序列进行卷积运算, 提取更多的特征数据, 相比传统卷积降低了模型的理论计算量. 实验结果表明, 该方法在越南语 4 种不同翻译粒度上均取得最佳效果, 一定程度上提升了汉越神经机器翻译性能.

关键词: 汉越神经机器翻译; 数据稀疏; 粒度; 深度可分离卷积

中图分类号: TP 391

文献标志码: A

文章编号: 0438-0479(2020)02-0220-05

近年来, 神经机器翻译^[1-2] (neural machine translation, NMT) 成为机器翻译研究的重要方向. 越南语属于资源稀缺型语言, 其特征是双语对齐语料贫乏, 语料收集困难, 使得数据稀疏问题十分严重, 极大地影响了模型的翻译效果. 在 NMT 中, 为了缓解数据稀疏的问题, 一般采用更小的输入/输出粒度来解决稀有和未知单词问题. Costajussà 等^[3] 以字符表示为底层, 采用卷积层和高速层神经网络形成词语表示, 实验中将所提出的模型应用于源语言, 以帮助编码器更完整地捕获源语言句子语义信息, 而目标语言仍以单词进行概率估计和生成; Chung 等^[4] 提出一种基于字母的编码器, 在目标语言端缓解了词汇量受限的问题; Ling 等^[5] 进一步在双语两端采用字符表示方法, 以字符表示为底层, 采用双向长短时记忆 (LSTM) 神经网络形成词语表示; Sennrich 等^[6] 采用字节对编码 (byte pair encoding, BPE) 算法对子词进行建模, 完成对单词的拆分; Luong 等^[7] 提出一种混合字符-词语模型, 其中词语级别模型负责词语表中词语的建模和生成工作, 字符级别模型负责源语言未知 (UNK) 符号所对应的单词建模工作和目标语言 UNK 符号的单词恢复工作.

在汉越神经机器翻译模型中, 可以对越南语这种

拼音文字进行切分, 使用切分后的语料进行模型训练, 以降低翻译中的数据稀疏问题, 这种方法已在英、法等西方拼音文字的翻译问题中被证明是十分有效的. 越南语可以切分为多种不同粒度, 除了词、字符等常见粒度切分外, 越南语还存在音节粒度, 每一个音节常常是一个有意义的单位, 可以独立使用, 这些单位又可作为构成多音节词的基础, 如同汉语拼音一样, 每一个音节可视为一个独立的“字”. 根据不同粒度切分后的序列中包含了许多局部特征信息, 这些信息对于 NMT 而言是十分有用的, 而卷积神经网络的特点就是能够提取局部特征向量, 同时还能保留特征之间的相对位置. 相比传统卷积, 深度可分离卷积是一种执行卷积运算更高效的方法, 需要的参数更少, 计算量也更小, 往往可以使用更少的数据学到更好的表示, 从而得到性能更好的模型.

本文提出一种基于深度可分离卷积的汉越 NMT 方法. 该方法首先根据越南语的语言特点, 将越南语切分为词、音节、字符、子词 4 种不同粒度, 通过切分降低低频词的数量. 其次利用深度可分离卷积改进 Vaswani 等^[8] 提出的基于注意力机制的 NMT 模型, 通过增加深度可分离卷积神经网络^[9-10], 对模型输入

收稿日期: 2019-08-30 录用日期: 2020-01-03

基金项目: 国家重点研发计划 (219QY1800, 2019QY1801); 国家自然科学基金 (61732005, 61672271, 61761026, 61762056, 61866020)

* 通信作者: 405904235@qq.com

引文格式: 徐毓, 赖华, 余正涛, 等. 基于深度可分离卷积的汉越神经机器翻译[J]. 厦门大学学报(自然科学版), 2020, 59(2): 220-224.

Citation: XU Y, LAI H, YU Z T, et al. Chinese-Vietnamese neural machine translation based on deep separable convolution[J]. J Xiamen Univ Nat Sci, 2020, 59(2): 220-224. (in Chinese)



的不同粒度序列进行卷积运算,优先学习序列的局部信息,提取更多的特征数据,同时减少卷积操作中所需参数数目,降低模型的理论计算量,提升了模型的表述能力。

1 不同粒度切分

在 NMT 任务中,对语料进行切分是预处理过程中非常重要的一个步骤. 语料中一个句子所包含的特征是由多个局部特征共同组成. 在语料预处理阶段,句子切分的粒度越大,切分结果越能够保存更加完整的局部特征,但同时会加重数据稀疏的问题;句子切分的粒度越小,所包含局部特征越少,数据稀疏的问题却会得到一定程度的缓解. 尤其在双语资源相对匮乏的汉越机器翻译任务中,切分粒度的把控就显得十分重要。

越南语和汉语属于同一种孤立语言. 在汉语中,单词不用空格区分,句子包括一系列连续的字符(包括标点符号). 在越南语中,相邻拼写的单词(音节)用空格隔开,标点符号位于拼写单词之后. 同时越南语还是一种拼音文字,一些适用于西方文字的粒度切分同样适用于越南语. 结合以上语言特点,本研究对越南语选取了词、音节、字符、子词 4 种不同的粒度切分来进行实验,各粒度切分示例如表 1 所示。

表 1 不同粒度切分示例

Tab. 1 Examples of segmentation with different granularity

粒度切分	示例
词	Bé, nhìn, vào, chúng tôi
音节	Bé nhìn vào chúng tôi
子词	B@@ é nhìn vào chúng tôi
字符	B é h ì n v à o c ú g t ô i

1.1 词粒度

现有的大多数 NMT 都以词为基本翻译单元,基于词的研究取得了全方位的发展,分词模型也变得更加简单有效,准确性很高. 本研究采用已有的越南语分词工具包对越南语进行分词处理。

1.2 音节粒度

越南语音节包括声母、韵母和声调 3 个部分. 声母由 23 个辅音承担,韵母由介音(元音 u)、主音(元音)和尾音(辅音韵尾)组成. 声调包括平声(Thanh ngang)、锐声(Thanh sắc)、玄声(Thanh huyền)、问声

(Thanh hỏi)、跌声(Thanh ngã)、重声(Thanh nặng)6 个声调,声调始终标识于音节中主音(元音)的上下。

越南语一共有 11 个单元音、23 个双元音、12 个三元音、25 个辅音. 如表 2 所示,根据韵母部分中元音的数量来称呼韵母类型为:单元音、二合元音、三合元音、单元音后附辅音、二合元音后附辅音、三合元音后附辅音,再由这些元音、辅音、声调组成音节,进而构成词。

表 2 越南语音节类型表

Tab. 2 Vietnamese syllable type table

示例	韵母名称
<u>bố</u> , <u>mẹ</u>	单元音
<u>cũa</u> , <u>tòa</u>	二合元音
<u>rượu</u> , <u>veau</u>	三合元音
<u>hát</u> , <u>phòng</u>	单元音后附辅音
<u>viết</u> , <u>tiếng</u>	二合元音后附辅音
<u>huyền</u> , <u>tuyết</u>	三合元音后附辅音

越南语中音节组合数量众多,大约有 2 500 个音节,但书写时使用空格隔开每一个音节,因此本研究利用空格对越南语音节进行切分处理,处理好的越南语音节就类似于汉语中的一个“字”。

1.3 子词粒度

BPE^[6]算法通过迭代地用一个未使用的字节替换序列中最频繁的字节对的过程. 其基本思路是:首先将语料以最小单元(越南语中指越南语字母,汉语中指汉字)切分,然后统计语料中所有相邻最小单元组合出现的频数,再找出频数最大的组合,将最大组合合并后替换语料库中原来组合,循环上面的操作. 越南语字母种类相对较少,因此切分后的语料与原文比较变化并不明显。

BPE 算法有两种应用方法:一是独立 BPE,即构建源语言词典和目标语言词典这两个独立词典;二是两种语言共同生成一个词典的联合 BPE. 理论上后者效果好一点,可以保证源语言和目标语言分割的一致性. 但是考虑到汉语和越南语并不属于拥有共享字母表的两种语言,所以本文采用独立 BPE 的方法,分别生成汉语词典和越南语词典。

1.4 字符粒度

通常语言学上把词作为基本意义单位,因此在自然语言处理中常使用词作为基本的翻译单元,但使用字符作为翻译的基本单元有其独特的优势. 字符可以加深语

言间的层次结构,将词和句子的两级层次结构加深到字符、词和句子的 3 级层次结构. 另外,使用字符切分粒度可以解决翻译中存在 UNK 符号的问题. 越南语构词复杂,包括单音节词、复合词、重音叠韵词、偶合词、派生词等. NMT 模型无法覆盖训练所有的词,所以 UNK 符号的问题一直存在于词级翻译模型中,但基本的越南语字符数量有限,而且任意越南语的词都是由字符序列组合而成. 将字符粒度用于汉越 NMT 中,字符间的组合规则都由 NMT 模型自动学习,为翻译模型学习这种语言间的组合关系提供了经验数据.

2 基于深度可分离卷积的 NMT

2.1 深度可分离卷积神经网络

深度可分离卷积^[9-10]将传统的卷积分解为一个深度卷积和一个逐点卷积,结构如图 1 所示. 深度卷积在输入的每个通道上独立地执行空间卷积,一个卷积核负责一个通道,一个通道只被一个卷积核卷积,因此一个深度卷积操作比传统卷积操作减少了不少的参数数量. 同时因为深度卷积对每一个通道都进行了学习,而不是所有通道对应同一个滤波器,所以可得到质量更佳的特征. 逐点卷积采用 1×1 的卷积窗口进行卷积,从而把深度卷积的输出映射到一个新的通道空间中.

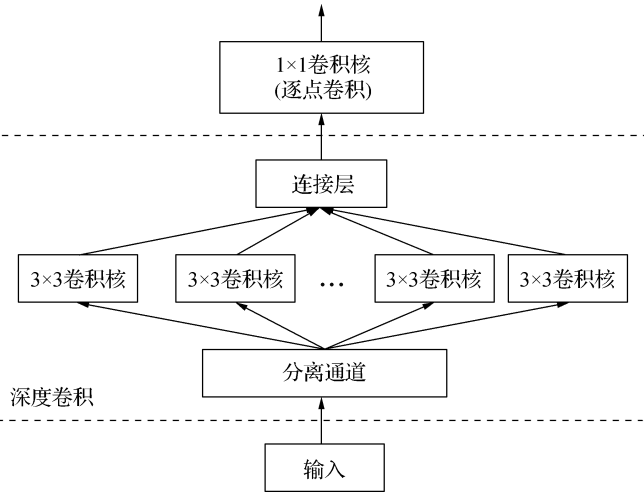


图 1 深度可分离卷积神经网络

Fig. 1 Deep separable convolutional neural network

传统卷积的计算量为 $D_K \times D_K \times M \times N \times D_F \times D_F$, 其中, D_K 为卷积核的大小, M 为输入的通道个数, N 为卷积核的个数, D_F 为输入的大小. 深度可分离卷积的计算量为 $D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F$. 深度可分离卷积与传统卷积的计算量比值

为 $N^{-1} + D_K^{-2}$. 由此可见,深度可分离卷积减少了大量的计算量.

常规的卷积层中,需要同时进行特征抽取和特征融合的工作,参数的使用效率低且效果不理想. 相反,深度可分离卷积把两步分离开来,先进行特征抽取,再进行特征融合,这样做可以充分利用模型参数进行表示学习,使用更少的参数,取得更好的效果.

2.2 基于深度可分离卷积的 NMT 模型

本研究以 Vaswani 等^[8]提出的基于注意力机制的 NMT 模型为基础模型,不同的是:本研究在输入层和编码器之间加入了深度可分离卷积神经网络,输入数据首先经过卷积运算之后再通过编码器进行其他运算. 编码器由 N_e 个相同的层堆叠在一起,每一层又包括两个子层,第一个子层是自注意力层,第二个子层是一个前馈神经网络层,自注意力层和前馈网络的输出分别被送到两个“求和与归一化层”进行处理,该层包含一个残差连接^[11]和一个层标准化^[12]步骤. 解码器采用同样的架构,只是在自注意力层和前馈神经网络层之间加入了一个注意力编-解码层,用来帮助解码器注意力输入句子的其他部分,这些子层同样也使用了残差连接和层标准化步骤. 模型编码器结构如图 2 所示.

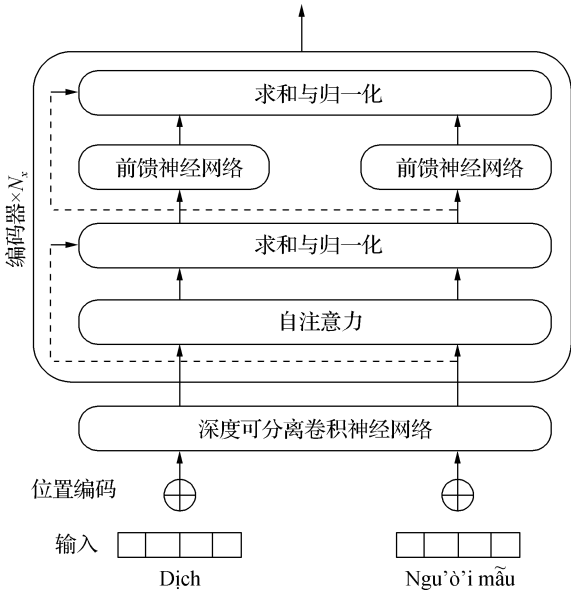


图 2 基于深度可分离卷积的 NMT 编码器

Fig. 2 Encoder of deep separable convolution based NMT

模型所有的子层以及嵌入层的输出维度都是 d_{model} , 本研究对模型的每一个输入都添加了位置向量. 模型计算自注意力的第一步是从每个编码器的输入向量上创建一个 Query 向量,一个 key 向量和一个

Value 向量^[10],这三个向量是通过词嵌入向量与训练过程中创建的训练矩阵产生的。

3 实验结果与分析

3.1 实验数据

实验中训练的汉越双语语料规模为 12 万句对,其中测试集 1 500 对双语句对,验证集 1 200 对双语句对。在训练之前对双语语料做了人工校对处理,去除语料中存在的重复、空格和不规则符号。

3.2 实验设置

实验中汉语仅使用结巴分词工具做分词处理,越南语按上述 4 种粒度切分。实验分为汉语-越南语(汉越)、越南语-汉语(越汉)两个翻译方向。选取目前效果较好的两个深度学习翻译框架——卷积神经网络(CNN)^[13-14]和 Transformer 与本文模型相比较。

实验中使用的词表大小为 32 000。CNN 模型由 5 层编码器(每层核宽度为 3)和 9 层解码器(每层核宽度为 3,每一层均配备一个注意力机制)构成,批大小为 128。Transformer 模型编码器和解码器层数为 6 层,词向量和隐层单元数为 512,批大小为 1 024。基于深度可分离卷积的汉越神经机器翻译模型卷积核大小为 3×3,每层核宽度为 3,编码器和解码器层数为 6,批大小为 1 024。所有实验均使用双语互译评估(BLEU)值作为翻译效果的评测指标。

3.3 实验结果

表 3 为不同翻译模型采用不同越南语切分粒度进行翻译时的 BLEU 值对比。可以看出:1) 本文模型相对于 CNN 与 Transformer 在 4 种不同粒度下的越南语切分方法均取得了最佳效果。例如,在子词粒度下,相比于 CNN,本文模型的 BLEU 值在汉越和越汉翻译方向上分类提升了 4.25 和 4.16 个百分点;相比于 Transformer,其 BLEU 值在汉越和越汉翻译方向上分别提升了 1.62 和 1.48 个百分点。在字符粒度下,相比 CNN,本文模型其 BLEU 值在汉越和越汉翻译方向上分别提升了 4.29 和 4.05 个百分点;相比 Transformer,其 BLEU 值在汉越和越汉翻译方向上分别提升了1.70 个和 1.64 个百分点。2) 无论是 CNN 和 Transformer,还是本文方法,汉越翻译方向的翻译效果都要优于越汉方向的翻译效果,BLEU 值平均约高出 0.88 个百分点。分析可能的原因为汉语构词复杂,在计算能力有限的情况下训练过程中涉及到的词数量有限,造成 BLEU 值偏低。3) 在所有模型中子词

粒度的效果最好,这是因为子词粒度切分可以将低频词切分成高频词片段,缓解语料匮乏带来的数据稀疏问题。而在所有模型中字符粒度效果较差的原因是越南语句子切分为字符后,虽然降低了数据稀疏问题,但是句子长度大大增加,增加了长距离依赖学习的难度。

本文还将传统卷积神经网络模型和深度可分离卷积神经网络模型的训练时间进行了对比。二者在 12 万的汉越双语语料上,分别约为 18 和 16 h,相比之下,深度可分离卷积神经网络模型的训练时间减少了 11%,在一定程度上提高了模型的计算效率。然而,在实际的训练过程中,模型的收敛速度取决于多重因素,因此该数据只能作为本文实验条件下的粗略估计。

表 3 不同模型中越南语切分粒度对 BLEU 值的影响
Tab. 3 BLEU values of Vietnamese with different granularities in different model

越南语 粒度	BLEU/%					
	CNN		Transformer		本文模型	
	汉越	越汉	汉越	越汉	汉越	越汉
词	17.45	16.77	19.63	18.76	21.24	20.56
音节	17.68	17.05	20.14	19.20	21.92	21.01
子词	18.32	17.38	20.95	20.06	22.57	21.54
字符	15.61	14.73	18.20	17.14	19.90	18.78

4 结 论

本文提出一种基于深度可分离卷积层的汉越神经机器翻译方法。相比传统卷积,深度可分离卷积模型训练时间减少了 11%。更高效的方法,需要的参数更少,计算量也更小,在越南语 4 种不同粒度上的实验结果表明,本文方法显著提高了越南语的翻译效果。另外,本文方法中所使用的汉越双语语料规模偏小,在未来的工作中考虑利用大规模单语数据来进一步提升汉越 NMT 的性能。

参考文献:

[1] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]// Advances in Neural Information Processing Systems, New York: MIT, 2014:3104-3112.

- [2] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]// International Conference on Learning Representations (ICLR). Cambridge: MIT Press, 2015: 1604-1614.
- [3] COSTAJUSSÀ M R, FONOLLOSA J A R. Character-based neural machine translation [C] // 54th Annual Meeting of the Association for Computational Linguistics: Short Papers, Stroudsburg: ACL, 2016: 357-361.
- [4] CHUNG J, CHO K, BENGIO Y. A character-level decoder without explicit segmentation for neural machine translation[C]// 4th Annual Meeting of the Association for Computational Linguistics: Long Papers, Stroudsburg: ACL, 2016: 1693-1703.
- [5] LING W, TRANCOSO I, DYER C, et al. Character-based neural machine translation [EB/OL]. [2019-08-01]. <https://arxiv.org/pdf/1511.04586.pdf>.
- [6] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[C]// 54th Annual Meeting of the Association for Computational Linguistics: Long Papers, Stroudsburg: ACL, 2016: 1715-1725.
- [7] LUONG M T, MANNING C D. Achieving open vocabulary neural machine translation with hybrid word-character models[C]// 54th Annual Meeting of the Association for Computational Linguistics: Long Papers, Stroudsburg: ACL, 2016: 1054-1063.
- [8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C] // Advances in Neural Information Processing Systems. New York: MIT, 2017: 5999-6009.
- [9] KAISER L, GOMEZ A N, CHOLLET F. Depthwise separable convolutions for neural machine translation[EB/OL]. [2019-06-15]. <https://arxiv.org/pdf/1706.03059.pdf>.
- [10] SLFRE L, MALLAT S. Rigid-motion scattering for texture classification[J]. Computer Science, 2014, 3559: 501-515.
- [11] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C] // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Los Alamitos: IEEE, 2016: 770-778.
- [12] BA L J, KIROS J R, HINTON G E. Layer normalization [EB/OL]. [2019-07-21]. <https://arxiv.org/pdf/1607.06450.pdf>.
- [13] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C] // 34th International Conference on Machine Learning, New York: IMLS, 2017: 1243-1252.
- [14] KALCHBRENNER N, ESPEHOLT L, SIMONYAN K, et al. Neural machine translation in linear time [EB/OL]. [2019-10-31]. <https://arxiv.org/pdf/1610.10099.pdf>.

Chinese-Vietnamese neural machine translation based on deep separable convolution

XU Yu, LAI Hua*, YU Zhengtao, GAO Shengxiang, WEN Yonghua

(Yunnan Key Laboratory of Artificial Intelligence, College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: In the Chinese-Vietnamese neural machine translation, the lack of Chinese-Vietnamese parallel corpus has created a serious problem of sparse data, which greatly affects the translation performance of the model. In order to improve the performance of Chinese-Vietnamese neural machine translation with sparse data, a deep separable convolution based Chinese-Vietnamese neural machine translation method is proposed. First, according to the linguistic characteristics of Vietnamese, the method divides Vietnamese into four different granularities: words, syllables, characters and subwords. Second, the neural machine translation model is improved by using the depth separable convolution. By adding depth separable convolution in traditional neural networks, more features can thus be extracted with convolution operation performed on sequence inputs of different granularities. Compared with the traditional convolution, the improved method reduces the theoretical calculation amount. Finally, experimental results show that the proposed method achieves considerably in four different translation granularities of Vietnamese, improving the neural machine translation ability in Chinese-Vietnamese to a certain extent.

Keywords: Chinese-Vietnamese neural machine translation; sparse data; granularity; depth separable convolution