文章编号：2095-2163(2021)04-0009-05 中图分类号：TP391.1 文献标志码：A

结合 AdaBERT 的 TextCNN 垃圾弹幕识别和过滤算法

孙瑞安，张云华
（浙江理工大学 信息学院，杭州 310018）

摘 要：为解决使用 BERT(Bidirectional Encoder Representations from Transformers)模型时，参数规模太大的问题，本文采用了结合 AdaBERT(Task-Adaptive BERT)的 TextCNN 算法。首先使用 AdaBERT 对弹幕文本进行学习，以更少的时间获得更有效的词向量；使用其生成的词向量作为 TextCNN 的输入；然后使用批量标准化，减少梯度消失的情况发生；最后使用 Softmax 进行分类概率计算。为了验证本算法的有效性，在弹幕数据集上进行训练，和多个文本分类算法进行对比实验。其结果表明，本算法可以改进算法运行速度，提高在垃圾弹幕识别和过滤上的性能。

关键词：AdaBERT；TextCNN；弹幕；文本过滤

TextCNN Based On AdaBERT Barrage Recognition and Filtering Algorithm

SUN Ruian, ZHANG Yunhua
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

【Abstract】 In order to solve the problem of too large parameter scale when using the BERT (Bidirectional Encoder Representations from Transformers) model, this article uses the TextCNN algorithm combined with AdaBERT (Task-Adaptive BERT). First use AdaBERT to learn the barrage text to obtain more effective word vectors in less time; then use the word vectors generated by it as the input of TextCNN; then use batch normalization to reduce the occurrence of gradient disappearance; finally use Softmax for Classification probability calculation. In order to show the goodness of this algorithm, training is conducted on the barrage data set, and comparison experiments are performed with multiple text classification algorithms. The results show that this model can enhance the running speed of the algorithm, and improve the performance of garbage barrage recognition and filtering.

【Key words】 AdaBERT; TextCNN; barrage; Text filtering

0 引 言

随着现代互联网的发展，越来越多的人在网络上寻找消磨时间的娱乐方式，其中就包括了带有趣味的弹幕视频——有弹幕飘过的视频。弹幕最初出现在日本 niconcio 视频网的视频里。之所以叫弹幕，是因为其就像子弹一样密集地在视频上飘过，网友借此发明视频弹幕这一网络词汇。人们可以使用弹幕发表对某一情节的看法和评论，也可以借用弹幕对一些电影进行背景介绍，让新来的观众对电影有一定的了解，方便决定自己是否要继续看下去。而有些视频的语音是外语的，而且没有提供字幕，这时候就有热心网友使用底部弹幕的形式制作中文字幕方便别人的观看。可以看出，弹幕作为一种新型网络文化有一定的趣味性和实用性。但是，当有人利用弹幕发布与视频无关的信息，比如广告、贬低别人的话语，又或者发布遮挡字幕的底部弹幕，就会影响他人的观看，甚至形成不好的社会风气，造成恶劣的后果。所以，对垃圾弹幕进行过滤是一件急需落实的措施。目前的弹幕过滤方法一般是使用关键词进行识别过滤。该方法将弹幕评论与关键词进行对比，如果匹配成功，则屏蔽该弹幕；否则不屏蔽[1]。在使用关键词进行垃圾弹幕过滤时，需要与时俱进更新新的屏蔽词，无形中又增加了时间及人力成本。所以，只使用关键词进行过滤，不仅效率较低，其准确率也不高。为了提高垃圾弹幕的识别和过滤效率，本文提出了一种结合 AdaBERT 自适应结构的 TextCNN 垃圾弹幕识别和过滤算法。与原始的 BERT 模型相比，使用 AdaBERT 压缩后的模型的参数规模大大下降，其推理速度也提升了十多倍，提升了垃圾弹幕识别模型的性能和效率。

1 相关研究

自然语言处理是机器学习的一个重要研究领域，而文本分类和文本生成是其两个研究重点。本文研究的弹幕就是一种特殊的网络文本。现在机器学习和深度学习在文本分类领域的研究发展迅速，并取得了一定的进展。

作者简介：孙瑞安(1995-)，男，硕士研究生，主要研究方向：自然语言处理；张云华(1965-)，男，博士，教授，主要研究方向：软件工程。

收稿日期：2020-12-12

机器学习的方法主要有 4 种:

(1)逻辑回归方法。这种方法经常用来预测一个样例属于某个类别的概率,适用于二分类问题和多分类问题;

(2)朴素贝叶斯方法[2]。其原理依赖于数理统计的贝叶斯定理;

(3)随机森林方法。这种方法是将多个决策树的结果综合起来[3];

(4)支持向量机(Support Vector Machine,SVM)方法。其可以用于线性分类、非线性分类、回归等任务,主要思想是使用间隔进行分类[4]。

随着深度学习的发展,越来越多的深度学习模型被应用于短文本分类任务中。如,文献[5]中提出基于自编码网络的短文本流形表示方法,实现文本特征的非线性降维,可以更好地以非稀疏形式、更准确地描述短文本特征信息,提高分类效率;文献[6]提出一种基于语义理解的多元特征融合中文文本分类模型,通过嵌入层的各个通路,提取不同层次的文本特征,比神经网络模型(Convolutional Neural Network,CNN)与长短期记忆网络模型(Long Short-Term Memory,LSTM)的文本分类精度提升了 8%;文献[7]使用 CNN 模型,将句中的词向量合成为句子向量,并作为特征训练多标签分类器完成分类任务,取得了较好的分类效果;文献[8]提出 DCNN 模型,在不依赖句法解析树的条件下,利用动态 k-max pooling 提取全局特征,取得了良好的分类效果;文献[9]采用多通道卷积神经网络模型进行监督学习,将词矢量作为输入特征,可以在不同大小的窗口内进行语义合成操作,完成文本分类任务;文献[10]结合 CNN 和 LSTM 模型的特点,提出了卷积记忆神经网络模型(Convolutional Memory Neural Network,CMNN),相比传统方法,该模型避免了具体任务的特征工程设计;文献[11]将 CNN 与循环神经网络(Recurrent Neural Network,RNN)有机结合,从语义层面对句子进行分类,取得良好的分类效果;文献[12]提出一种基于注意力机制的卷积神经网络,并将该网络用在句子对建模任务中,证明了注意力机制和 CNN 结合的有效性;文献[13]提出了一种基于弹幕内容和发送弹幕的用户标识的混合垃圾弹幕识别过滤算法,其主要考虑弹幕本身的特点来研究。

目前,迁移学习在自然语言处理的应用,主要针对第一层的微调预训练的词嵌入,而且对于不同的语言任务都要有针对性地单独训练一个模型,比较浪费时间和资源。为此,一些学者提出,通过一个大数据集下训练过的 NLP 模型,然后针对不同的小任务只需要细微的调些参数即可完成不同的语言处理任务。在这其中就包括 BERT 预训练模型[14]。BERT 模型是谷歌在 2018 年提出的,其在 11 个 NLP 任务中打败其它所有选手,成为最受瞩目的明日之星。BERT 使用 Transformer 进行特征提取,Transformer 可以学习到语句的双向关系。BERT 主要使用 MLM(Mask Language Model)和 NSP(Next Sentence Predication)作为训练任务。使用 BERT 预训练的模型只需进行微调参数就能适应各种下游任务,但 BERT 所需调整的参数量十分巨大,需要更好的硬件条件来运行。如何压缩 BERT 模型就成为某些研究者新的研究课题。Chen[15] 等通过可微神经架构搜索(Differentiable Neural Architecture Search, DNAS)将 BERT 压缩成适应相应任务的微小模型,加快了推理速度,减少了大量参数。

2 结合 AdaBERT 的 TextCNN 垃圾弹幕识别及过滤算法模型构建

2.1 AdaBERT 词向量模型

在使用 BERT 预训练模型时,其参数达到 110M 之多,给训练此模型带来一定难度。要想训练这样规模的模型需要更好的机器和更多的资费,这对于一般人是无法承担的。为在模型结构不变的情况下减少参数的规模,文献[16]提出向量参数分解的方法,将词语的大向量分解为小向量,并且将层之间的参数共享,实现了模型压缩。由于这些研究都是在不改变原始模型结构的情况下减少参数数量,而 BERT 在海量数据中学到了不同领域的知识。对于不同的任务,知识面是不同的。因此,需要寻找适合每种任务本身的、小的结构和知识。而 AdaBERT 就实现了这一目标。

AdaBERT 的损失函数包含两个方面:一个是针对任务进行知识蒸馏,引导模型的搜索;二是模型效率反馈的损失,对模型的搜索过程进行剪枝。只有同时考虑这两方面的损失,才不会导致最终的模型只有效率高而有效性低,或者只有有效性高而速度却很慢。而应该找到一个效率和有效性权衡的模型。具体流程如图 1 所示。

图中,目标弹幕文本数据集为 D_t,经过调整参数后的 BERT 模型记为 $BERT_t$,所探索的模型空间记为 S,而最终搜索到的最适合本文文本类型的模型记为 $s \in S$。其损失函数为:

$$L = (1-\lambda)L_c(s,\omega_s,D_t) + \lambda L_k(s,\omega_s,BERT_t) + \alpha L_e(s). \quad (1)$$

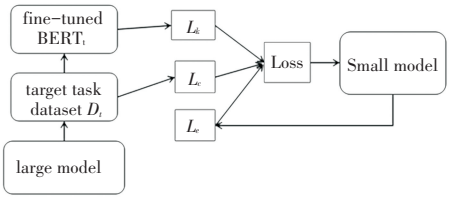

图 1　AdaBERT 流程

Fig. 1　AdaBERT process

式中，ω_s 是搜索到的结构 s 对应的训练权重；L_c 是和目标数据集 D_t 相关的交叉熵损失；而 L_k、L_e 分别是面向任务的知识蒸馏损失和模型的效率损失；λ 和 α 是平衡所有损失的超参数。

为了将搜索目标表示为分布变量，最直接的方式是建模为 one-hot 变量。但这样带来的问题是，离散的采样过程会使得梯度无法回传。因此，AdaBERT 引入了 Gumbel Softmax[17] 技术，将 one-hot 的模型结构变量松弛为连续分布 y^K 和 y^o。对于堆叠层数 K 相对应的第 i 维（表示模型结构最后堆叠 i 层的概率），以及候选 Opearation 的第 i 维（表示 DAG 中某条边最后导出第 i 种 operation 的概率）：

$$y^K = \frac{\exp[(\log(\theta_i^K) + g_i)/\tau]}{\sum_{j=1}^{K_{\max}} \exp[(\log(\theta_j^K) + g_j)/\tau]}, \tag{2}$$

$$y_i^o = \frac{\exp[(\log(\theta_i^o) + g_i)/\tau]}{\sum_{j=1}^{|O|} \exp[(\log(\theta_j^o) + g_j)/\tau]}. \tag{3}$$

这里，g_i 是 Gumbel 分布中采样得到的随机噪声，τ 代表此分布与 one-hot 分布的接近程度。此后，变量都是可微的，可以直接使用相应的优化器进行损失优化。

2.2　TextCNN 模型

使用 TextCNN 可以实现对文本的分类任务，其模型结构如图 2 所示。其中包括：一个用于生成词向量的嵌入层；一个包含几个卷积核的卷积层，一个卷积核可以得到 len(seq)-filter_size+1 个卷积结果；进入激活函数进行非线性化操作；再进行最大化池化操作；最后经过全连接传入 softmax 进行分类。

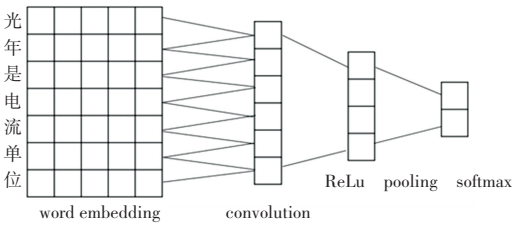

图 2　TextCNN 模型

Fig. 2　TextCNN model

2.3　结合 AdaBERT 的 TextCNN 模型

在 TextCNN 中一般使用 word2vec 或者 GloVe 作为词向量的选择，而 AdaBERT 使用 Transformer 可以真正识别上下文的信息。所以，本文使用 AdaBERT 的词向量代替 TextCNN 本身的词向量。BERT 模型本身学习了大量百科知识，拥有很好的学习能力来学习弹幕中的上下文关系。而 AdaBERT 可以训练出适合本文弹幕语料集的相应结构的模型，使用 AdaBERT 词向量，对提高最终的模型效率和有效性有一定作用。

结合 AdaBERT 的 TextCNN 模型，在输入层对文本使用 AdaBERT 转换成相应的词向量，然后将所有词向量拼接成一个向量矩阵 $\boldsymbol{B}$，公式如下：

$$\boldsymbol{B}_{1:n} = [b_1, b_2, \ldots, b_n]. \tag{4}$$

其中，[ ]代表拼接词向量的操作；b_i 代表句子里的每一个词向量；$B_{i:j}$ 代表将第 i 个词向量到第 j 个词向量拼接。然后使用不同的卷积核 W，大小（h）分别为 3、4、5。从而获得 3 个字符、4 个字符、5 个字符之间的关系。进行卷积操作得到特征 F_i，如式(5)所示：

$$F_i = \mathrm{Re}lu(W \cdot \boldsymbol{B}_{i:i+h-1} + b). \tag{5}$$

式中，b 为偏差，通过 *ReLU* 激活函数生成特征 $F = [F_1, F_2, \ldots, F_{n-m+h}]$，然后进行批量归一化（BN）操作，防止维度爆炸或者弥散。再进行最大池化，最后全连接到 softmax 层，输出样本在不同分类上的概率，取最大值为分类结果。

本文结合 AdaBERT 的 TextCNN 模型架构如图 3 所示：

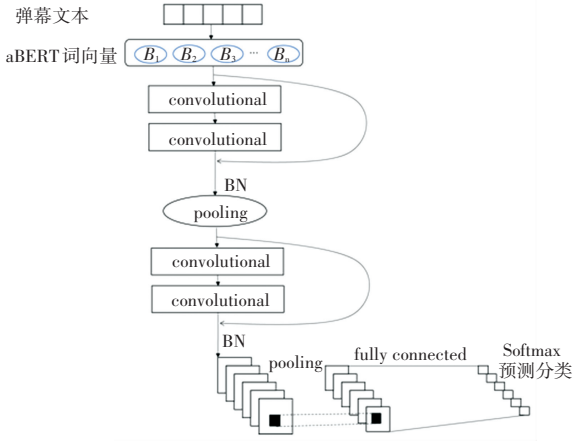

图 3　结合 AdaBERT 的 TextCNN 模型

Fig. 3　TextCNN based on AdaBERT model

3　实验及结果分析

3.1　数据集和实验环境

实验主要进行模型正确性的验证。本文采集了

哔哩哔哩弹幕网的《秒速五厘米》的弹幕数据作为实验的数据集。使用爬虫技术共爬取 5 384 000 条弹幕,经过去重、去除只有标点符号、去除表情等清理数据的手段后,剩余 154 268 条弹幕。对弹幕数据进行敏感词标注,再进行人工标注查漏补缺。由于垃圾弹幕属于少量异常数据,所以本文将取出与垃圾弹幕相等数量的正常弹幕,生成平衡数据集。

最终电影《秒速五厘米》的弹幕数据集一共包含 6 000 条带有标签的弹幕数据,其中含有 3 000 条正常弹幕和 3 000 条垃圾弹幕。弹幕数据集的标注结果见表 1。

表 1 弹幕数据集的标签

Tab. 1 Label of barrage dataset

弹幕内容	标签
山前没相见 山后别相逢	正常弹幕
明知从未开始,但仍心之向往	正常弹幕
初心总在变 哪有不变的道理 就看变的味道了	正常弹幕
我是第一吗	垃圾弹幕
光年是电流单位	垃圾弹幕

3.2 实验结果评价标准

评判分类问题性能优劣,一般可以用正确率和错误率来评估。而在本文的数据集中,少数异常是主要的关注对象,其分类精度也就显得很重要。数据集中正常弹幕和垃圾弹幕的数量差距大,是一种不平衡的文本分类数据集,那么正误率不太适合作为这种数据集的分类算法评判指标。本文将采用精确率(*Precision*)、召回率(*Recall*)、*F*1 分数(*F*1 - *score*)这 3 个指标对算法进行评估。表 2 所示的混淆矩阵更能直观地说明这 3 个概念。

表 2 混淆矩阵

Tab. 2 Confusion matrix

实际	预测	
	True	False
True	TP	FN
False	FP	TN

精确率 P 表示的是预测结果为正例的数据中预测正确的比例;召回率 R 是指实际为正例中预测为正例的百分比。精确率和召回率之间存在一定的数量关系,即当精确率上升时,召回率会下降,反之亦然。综合考虑精确率和召回率时可以使用 $F1$ 分数。以下是精确率、召回率和 $F1$ 分数的计算公式:

$$P = \frac{TP}{TP + FP}. \tag{6}$$

$$R = \frac{TP}{TP + FN}. \tag{7}$$

$$F1 - score = \frac{2}{\frac{1}{P} + \frac{1}{R}}. \tag{8}$$

3.3 实验过程与结果分析

本次实验使用 Windows10 操作系统、jupyter lab 平台,使用 TensorFlow 深度学习框架进行模型训练,主要开发语言为 Python。实验数据集包含相等数量的正常弹幕和垃圾弹幕,从中随机将数据集按 8:2 的比例分成训练集和测试集。为说明本文提出的结合 AdaBERT 的 TextCNN 算法的优势,通过与 TextCNN 算法、朴素贝叶斯算法和 BiLSTM 算法进行对比结果,说明本文算法的有效性。实验结果见表 3。

表 3 实验结果

Tab. 3 Experimental result

模型	指标		
	精确率	召回率	F1 分数
TextCNN	94.91%	96.55%	95.72%
Native-Bayes	87.98%	91.54%	89.72%
BiLSTM	91.09%	68.29%	78.06%
AdaBERT-TextCNN	98.78%	98.88%	98.83%

从测试结果可见,本文算法的 3 个指标都是最高的。与使用 word2vec 的其它算法相比,采用 AdaBERT 词向量模型的 TextCNN 模型相关指标均更高。说明使用 AdaBERT 进行模型预训练得到的词向量比 word2vec 词向量更好。使用统计学知识计算分类概率的朴素贝叶斯模型,没有考虑词之间的上下文关系,而只是把每个词单独转换成相应的数值,并计算其属于某个类型的概率。其取得的结果必定是不准确的。TextCNN、AdaBERT-TextCNN 模型都属于 CNN 类别的模型,而 BiLSTM 则属于 RNN 模型。CNN 类型的模型比使用 RNN 的 BiLSTM 的精确率和召回率更高,这说明弹幕这种短文本类型的分类更适合使用 CNN 进行。在垃圾弹幕识别中上下文关系比较少,关键词的信息更多。

为了说明 AdaBERT 对 BERT 的参数优化,本文还包含了这两种方法的实验对比,结果见表 4。

表 4 时间对比

Tab. 4 Time comparison

模型	精确率	召回率	F1 分数	时间/s
BERT-TextCNN	97.60%	96.71%	97.15%	15 821
AdaBERT-TextCNN	98.78%	98.88%	98.83%	11 144

从表中结果来看，使用自适应的 BERT 模型的确减少了训练时间，提高了模型的效率。

总体来看，本文提出的结合 AdaBERT 的 TextCNN 模型，在实验中取得较好的成果，与普通分类算法相比优势较大。使用 AdaBERT 相比一般 BERT 算法的参数更少，可以加快模型的预训练，更好的提取词向量特征，结合 TextCNN 后可以获得更好的模型泛化能力。可以预见，本文算法对垃圾弹幕过滤这一应用场景有较大作用，可以投入到实际的弹幕过滤系统中使用。

4　结束语

本文提出的结合 AdaBERT 的 TextCNN 垃圾弹幕识别与过滤算法模型，与以前的基于统计学的分类算法相比，有更高的准确率；与 CNN 类型和 RNN 类型的模型相比，拥有更好的泛化能力。采用 AdaBERT 也减少了模型的复杂程度，使得总体训练时间降低。实现了对垃圾弹幕文本更好的语义理解，获取了更准确的弹幕特征，提高垃圾弹幕识别的准确率。目前，本文只研究基于弹幕文本内容的筛选，后续将考虑加入弹幕的位置和视频内容等维度加以综合评估，进一步提高识别精确率。

参考文献

[1] 汪舸，吴方君. 基于种子词和数据集的垃圾弹幕屏蔽词典的自动构建[J]. 计算机工程与科学，2020，42(7)：1302-1308.

[2] Fredrik Ronquist, John P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models[J]. Bioinformatics, 2003, 19(12): 1572-1574.

[3] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45: 5-32.

[4] AMAYRI O, BOUGUILA N. A study of spam filtering using support vector machines[J]. 2010, 34(1): 73-108.

[5] 魏超，罗森林，张竞，等. 自编码网络短文本流形表示方法[J]. 浙江大学学报(工学版)，2015，49(8)：1591-1599.

[6] 谢金宝，侯永进，康守强，等. 基于语义理解注意力神经网络的多元特征融合中文文本分类[J]. 电子与信息学报，2018，40(5)：1258-1265.

[7] 孙松涛，何炎祥. 基于 CNN 特征空间的微博多标签情感分类[J]. 工程科学与技术，2017，49(3)：162-169.

[8] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv: 1404.2188, 2014.

[9] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv: 1408.5882, 2014.

[10] 郑啸，王义真，袁志祥，等. 基于卷积记忆神经网络的微博短文本情感分析[J]. 电子测量与仪器学报，2018，32(3)：195-200.

[11] HSU S T, MOON C, JONES P, et al. A hybrid CNN-RNN alignment model for phrase-aware sentence classification[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017: 443-449.

[12] Wenpeng Yin, Hinrich Schütze, Bing Xiang, et al. ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 259-272.

[13] 张树华. 基于内容和用户标识的混合型垃圾弹幕识别与过滤研究[D]. 杭州：杭州电子科技大学，2017.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[15] CHEN D, LI Y, QIU M, et al. Adabert: Task-adaptive bert compression with differentiable neural architecture search[J]. arXiv preprint arXiv: 2001.04246, 2020.

[16] LAN Z, CHEN M, GOODMAN S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv: 1909.11942, 2019.

[17] JANG E, GU S, POOLE B. Categorical Reparameterization with Gumbel-Softmax[C]// ICLR. 2017.

(上接第 8 页)

[15] SHAMIR A, TAUMAN Y. Improved Online/Offline Signature Schemes[C]//Annual International Cryptology Conference, 2001: 355-367.

[16] ZHANG L, WU Q, DOMINGO-FERRER J, et al. Distributed Aggregate Privacy-Preserving Authentication in VANETs[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(3): 516-526.

[17] VERHEUL E, HICKS C, GARCIA F D. IFAL: Issue First Activate Later Certificates for V2X[C]// 2019 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2019: 279-293.

[18] SATTAM S, RIYAMI A, PATERSON K G. Certificateless Public Key Cryptography[J]. Lecture Notes in Computer ence, 2003: 452-473.

[19] HORNG S J, TZENG S F, HUANG P H, et al. An efficient certificateless aggregate signature with conditional privacy-preserving for vehicular sensor networks[J]. Information ences An International Journal, 2015, 317: 48-66.

[20] MALHI A K, BATRA S. An efficient certificateless aggregate signature scheme for vehicular ad-hoc networks[J]. Discrete Mathematics & Theoretical Computer ence, 2015, 17(1).

[21] KERINS T, MARNANE W P, POPOVICI E M, et al. Efficient Hardware for the Tate pairing calculation in characteristic three[C]// International Workshop on Cryptographic Hardware and Embedded Systems. Springer, Berlin, Heidelberg, 2005: 412-426.

[22] BEUCHAT J L, DOI H, FUJITA K, et al. FPGA and ASIC implementations of the [formula omitted] pairing in characteristic three[J]. Computers & Electrical Engineering, 2010, 36(1): 73-87.