深度之眼
deepshare.net

本期论文主题:Bert

导师：Yamada

《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》

预训练的深度双向transformer用于语义理解

作者：Jacob Devlin

单位：Google

发表会议及时间：2018

深度之眼
deepshare.net

前期知识储备

Pre-knowledge reserve

概率论

了解基本的概率论知识，掌握条件概率的概念和公式

语言模型

掌握语言模型的原理，了解语言模型的评价标准

Transformer

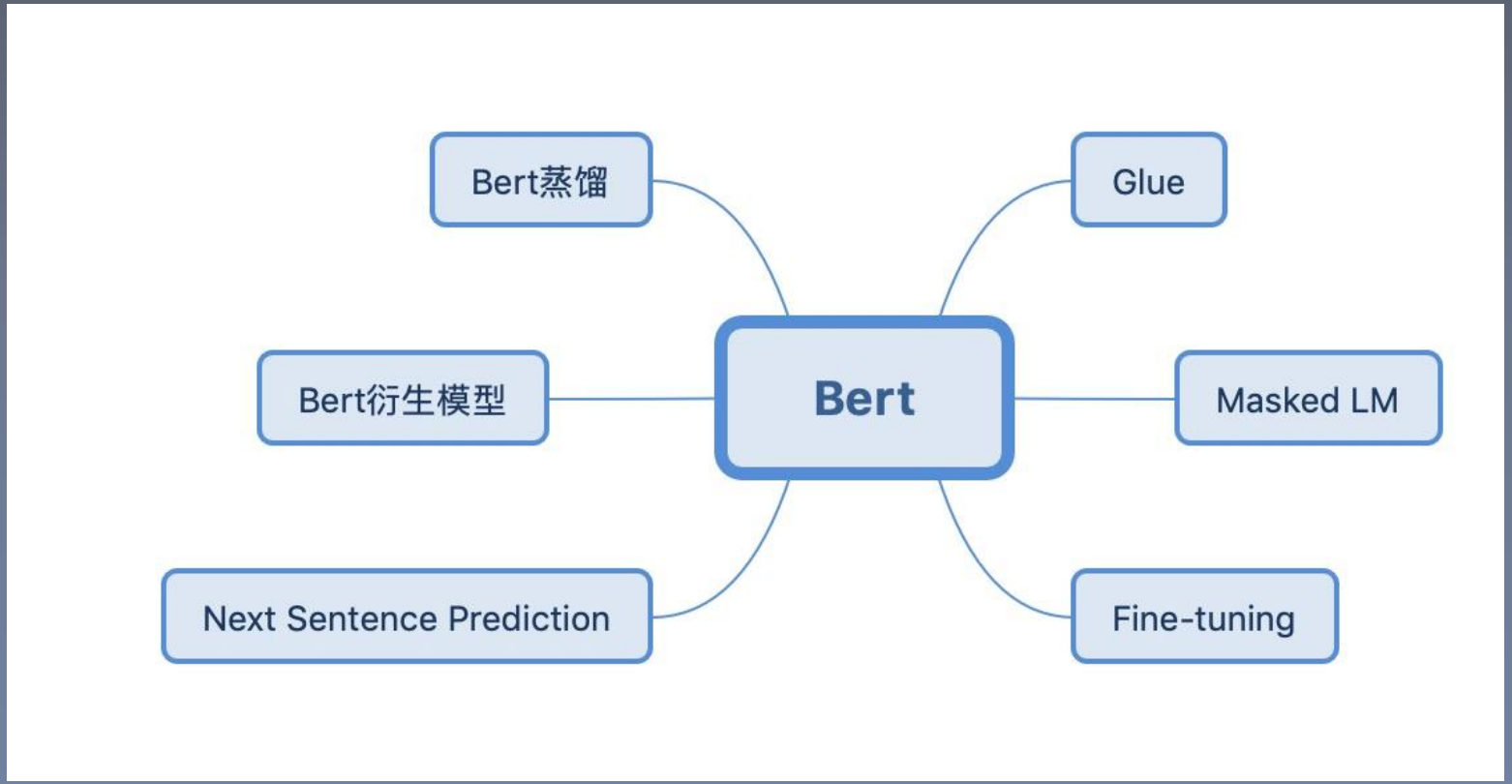
掌握Transformer的基本工作原理。

注意力机制

了解注意力机制的思想，掌握注意力机制的分类和实现方式

学习目标

Learning objectives



深度之眼
deepshare.net

课程安排

The schedule of course

01

Part 1
论文导读
背景介绍

02

Part2
Model Architecture

03

Part 3
Pre-training
Bert

04

Part 4
Fine-tuning Bert

05

Part 5
代码实现分析

深度之眼
deepshare.net

第一课：论文导读

The first lesson: the paper guide

深度之眼
deepshare.net

目 录

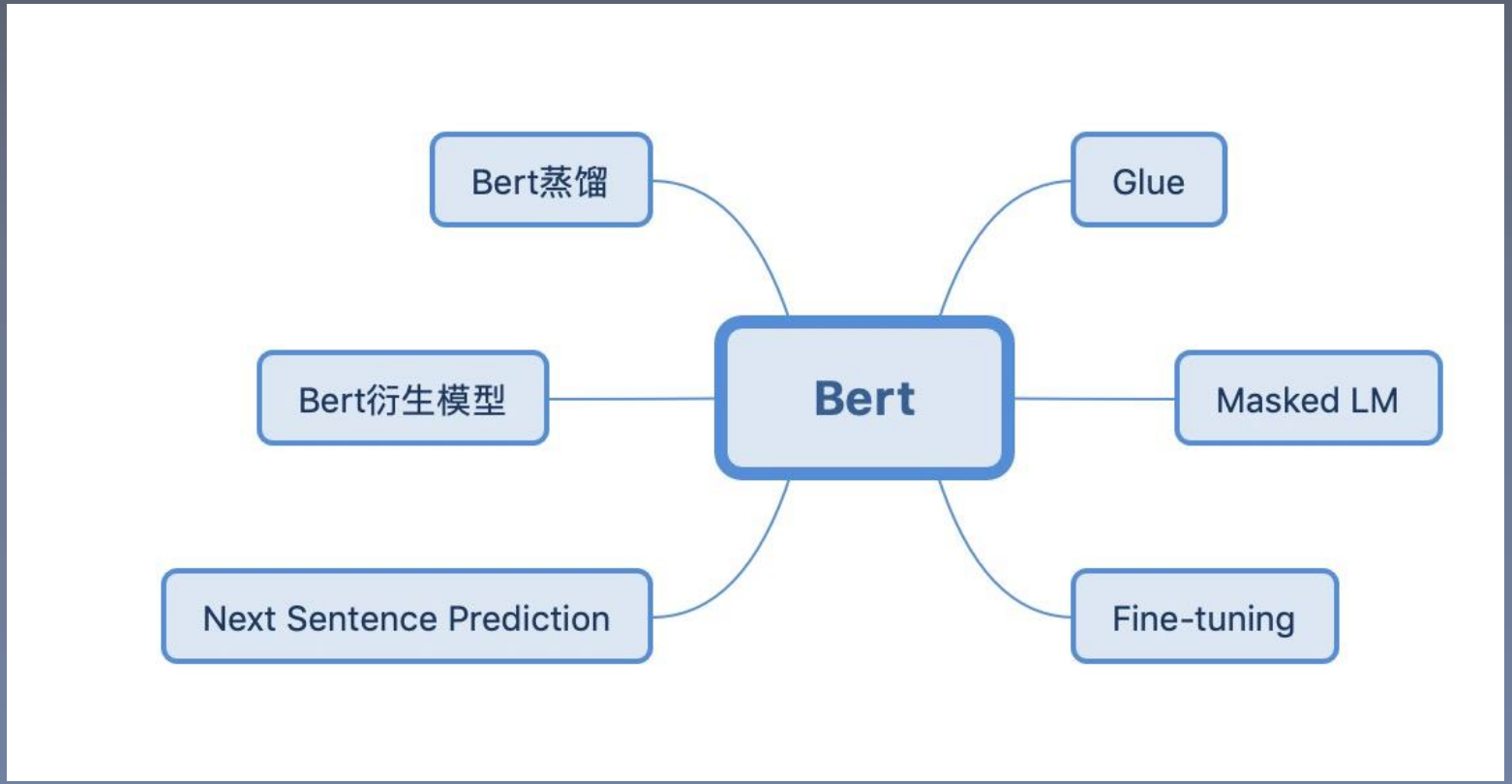
1 论文研究背景、成果及意义

2 论文泛读

3 Bert衍生模型 和Bert、Elmo、GPT比较

4 本课回顾及下节预告

知识树



论文研究背景、成果及意义

研究背景

Research background

深度之眼
deepshare.net

重点　重点来了！

Glue Benchmark

Table 1: A list of the different tasks and datasets used in our experiments.

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

Glue是用于衡量通用NLP模型的基准
https://gluebenchmark.com/leaderboard

研究背景

Research background

Feature-Based and Fine -tuning

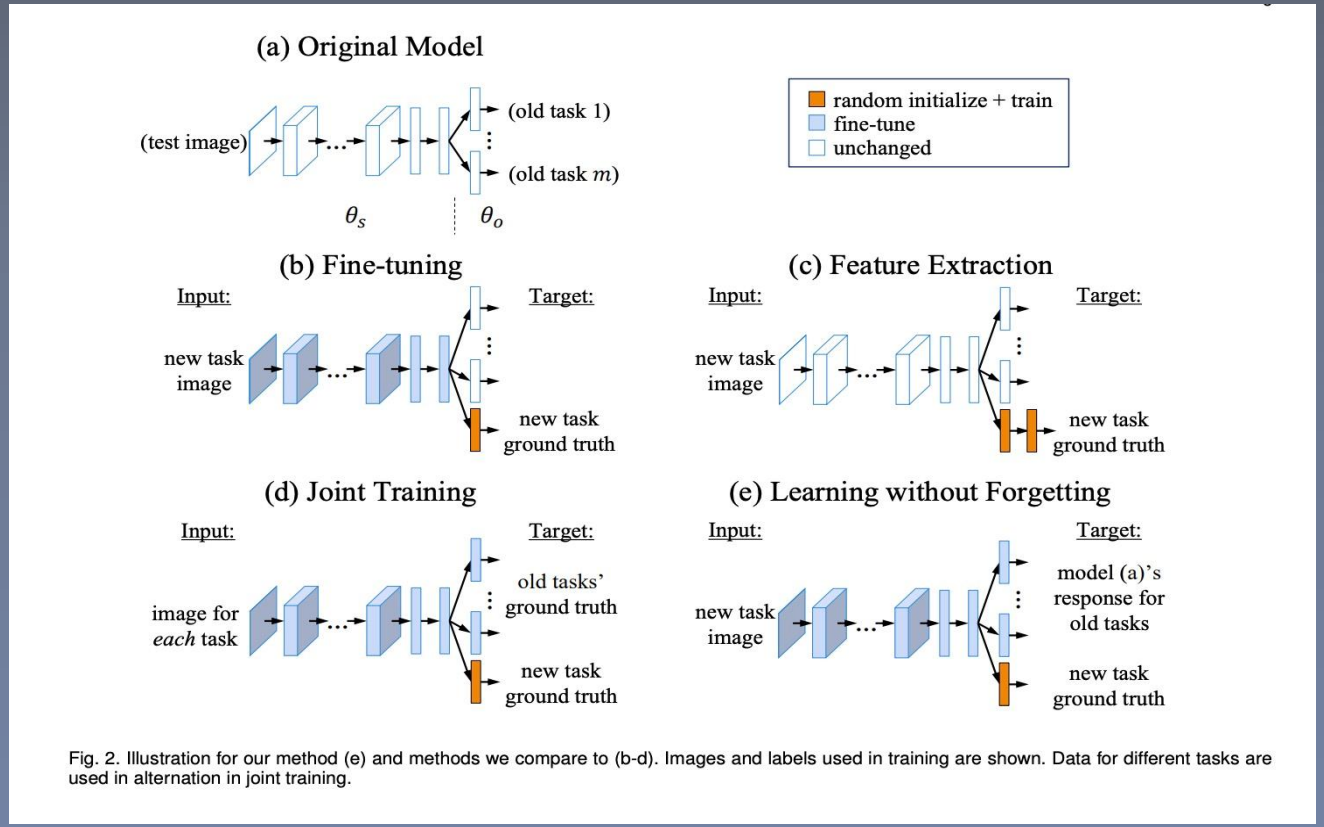


Fig. 2. Illustration for our method (e) and methods we compare to (b-d). Images and labels used in training are shown. Data for different tasks are used in alternation in joint training.

https://arxiv.org/pdf/1606.09282.pdf

研究成果

Research Results

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Bert在下游任务中的表现完全远超之前的模型。

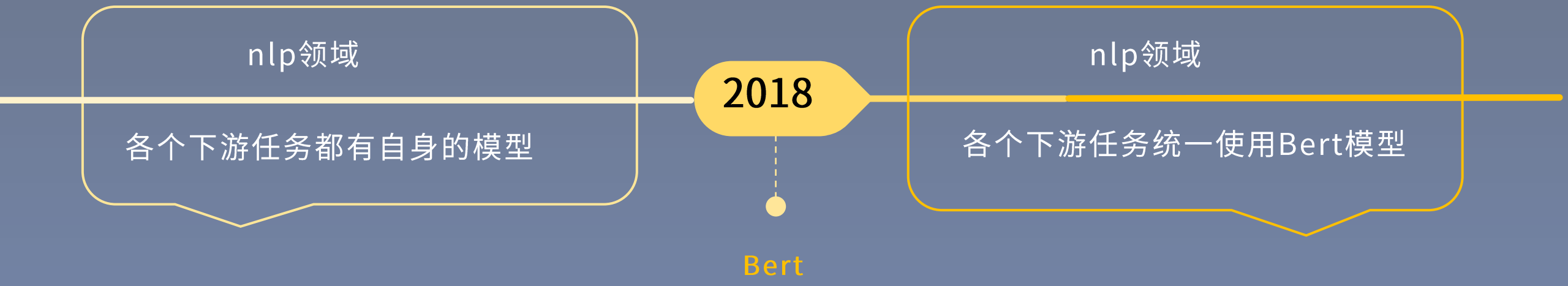
研究意义

Research Meaning

重点 重点来了！

Bert历史意义

- 获取了left-to-right和right-to-left的上下文信息。
- nlp领域正式开始pretraining+finetuning的模型训练方式



研究意义

Research Meaning

重点 重点来了！

Bert历史意义

- 获取了left-to-right和right-to-left的上下文信息。
- nlp领域正式开始pretraining+finetuning的模型训练方式

深度之眼
deepshare.net

论文泛读

Strcuture of Paper

论文结构

Structure of Papers

摘要 Abstruct

介绍背景及提出bert模型，在多个数据集上的效果都表现优异

1 Introduction

介绍之前的预训练模型存在单向的问题，因此提出Bert

2 Related Work

介绍无监督的fature-based和fine-tuning

3 The Architecture

Bert网络结构及其内部细节：pre-training和fine-tuning

4 Experiments

Bert在各个nlp下游任务中的结果对比。

5 Analysis

通过消融分析来验证Bert的pre-training和model size的有效性

6 Conclusions

论文验证了bert模型确实能在下游任务中有好的效果。

摘要
abstract

摘要核心

1. 我们提出了一种新的语言表征模型bert，不同于其他的语言表征模型，bert可以同时学习到向左和向右的上下文信息。

2. 预训练好的bert可以直接fine-tuning，只需加相应的输出层，无需太多模型结构的改动。

3. bert模型在各项nlp下游任务中都表现得良好。

论文小标题

Paper title

1. Introduction
2. Related Work
   2.1 Unsupervised Feature-based Approaches
   2.2 Unsupervised Fine-tuning Approaches
   2.3 Transfer Learning from Supervised Data
3. BERT
   3.1 Pre-training BERT
   3.2 Fine-tuning BERT
4. Experiments
   4.1 GLUE
   4.2 SQuAD v1.1
   4.3 SQuAD v2.0
   4.4 SWAG
5. Ablation Studies
   5.1 Effect of Pre-training Tasks
   5.2 Effect of Model Size
   5.3 Feature-based Approach with BERT
6. Conclusion

深度之眼
deepshare.net

Bert衍生模型以及Elmo、GPT、Bert对比

Strcuture of Paper

Bert衍生模型

Structure of Papers

衍生模型	模型特点	论文地址
RoBERTa	模型更大，参数量更多，静态mask变成动态mask	https://arxiv.org/pdf/1907.11692
ALBERT	参数量减少，跨层的参数共享	https://arxiv.org/pdf/1909.11942
BERT-WWM	全词mask，中文	https://arxiv.org/pdf/1906.08101
ERINE	mask实体，中文	https://arxiv.org/pdf/1904.09223v1
SpanBERT	随机选取span进行mask	https://arxiv.org/pdf/1907.10529
TinyBERT	对transformer结构进行蒸馏	https://arxiv.org/pdf/1909.10351
Sentence-BERT	孪生网络	https://arxiv.org/pdf/1908.10084
K-BERT	知识图谱	https://arxiv.org/pdf/1909.07606v1

Elmo、GPT、Bert比较

Structure of Papers

模型	模型采用结构	预训练形式	优点	缺点	在Glue上表现
ELMO	Bilstm+LM	featue-based	动态的词向量表征	双向只是单纯的concat两个lstm，并没有真正的双向	最差
GPT	Transformer Deocder部分(含有sequence mask，去掉中间的encoder-decoer的attention)	fine-tuning	在文本生成任务上表现出色 同时采用辅助目标函数和lm	单向的transformer结构，无法利用全局上下文信息	较差
BERT	Transformer Encoder部分(无sequence mask)	fine-tuning	在各项下游任务中表现最好 采用mlm的实现形式完成真正意义上的双向 增加了句子级别预测的任务	在文本生成任务上表现不好	最好

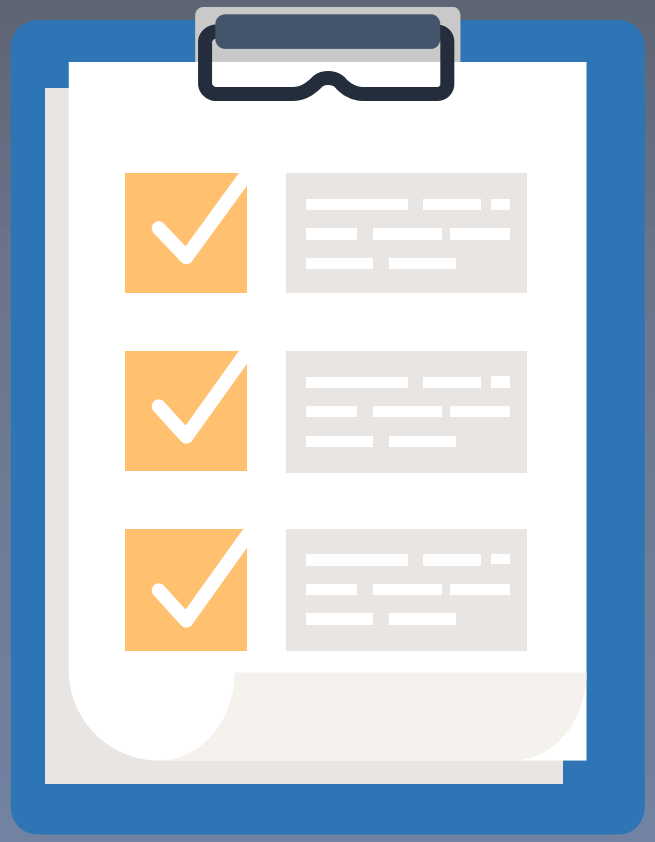
深度之眼
deepshare.net

本课回顾及下节预告

Review in the lesson and Preview of next lesson

本课回顾

Review in the lesson



深度之眼
deepshare.net

01 研究背景及成果意义

学习了GLUe以及概念feature-based和fine-tuning、了解了论文的实验结果。

02 论文总览

论文总共包含6个部分，论文主要介绍bert的结构。

03 Bert的衍生模型和Elmo、GPT、Bert的比较

学习了Bert的衍生模型，比较了Elmo、GPT以及Bert。

下节预告

Preview of next lesson

01 Pre-training Bert

学习Bert的pre-training部分

02 Fine-tuning Bert

学习Bert的fine-tuning部分

03 实验设置及结果分析

比较了模型在几个数据集上的表现情况。

04 论文总结

总结论文中创新点、关键点及启发点

下节课前准备

Preview of next lesson

- 下载论文
- 泛读论文
- 筛选出自己不懂的部分，带着问题进入下一课时

结 语

循循而进，欲速则不达也。

深度之眼
deepshare.net

联系我们：
电话：18001992849
邮箱：service@deepshare.net
ＱＱ：2677693114

公众号

客服微信