（装订线内不要答题）

复旦大学计算机学院/信息学院

2021~2022 学年第一学期期末考试试卷

■ A 卷 □ B 卷 □ C 卷

课程名称： 自然语言处理 **课程代码：** COMP130141.01/02

开课院系： 计算机科学技术学院/信息科学与工程学院 **考试形式：** 闭卷

姓名： _____ **学号：** _____ **专业：** _____

提示：请同学们秉持诚实守信宗旨，谨守考试纪律，摒弃考试作弊。学生如有违反学校考试纪律的行为，学校将按《复旦大学学生纪律处分条例》规定予以严肃处理。

题号	1	2	3	4					总分
得分									

（以下为试卷正文**或课程论文题目**）

一、 计算图（25%）

门控循环神经网络（GRU）是循环神经网络（RNN）的一个变体，可以有效地解决简单RNN的梯度消失问题，且相较于LSTM（长短期记忆网络）具有结构更简单，参数更少的优点。在序列标注、文本分类、机器翻译等任务中有着广泛的应用。以下给出 GRU 循环单元结构，请你基于该图完成以下问题：

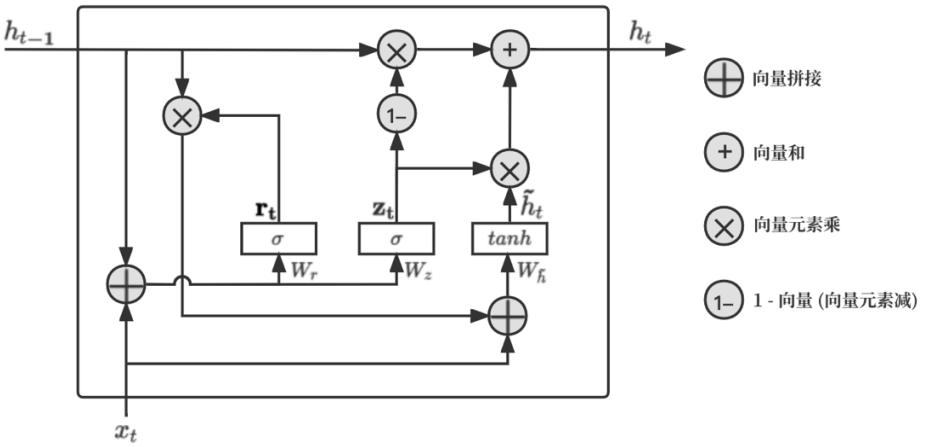


图1 GRU的循环单元结构

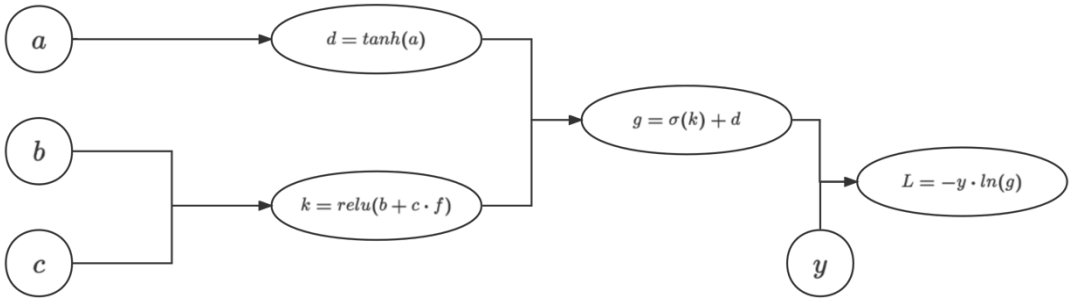
（1）请根据以上GRU循环单元的示意图写出由x_t，h_{t-1}计算得到h_t的运算式。

（提示：W_r，W_z，$W_{\tilde{h}}$对应的偏置分别为b_r，b_z，$b_{\tilde{h}}$，$\boldsymbol{r}_t$和$\boldsymbol{z}_t$是GRU的两个门，称为复位门和更新门，操作符$(1-$向量$)$会将1广播为向量的维度然后做向量元素减）

（2）根据以上GRU循环单元的计算过程，请根据自己的理解简要描述为什么GRU相较于简单RNN更善于捕捉序列的长距离依赖（不必进行反向传播严格公式推导证明）

（3）计算图是数学运算的图形化表示，通过建立神经网络的计算图，可以高效地执行网络参数的微分，基于此开发的深度学习框架可以实现自动梯度计算，从而大幅提高了开发效率。请根据以下计算图写出$\frac{\partial L}{\partial e}$，$\frac{\partial L}{\partial a}$的表达式。

（提示：$tanh(x) = \frac{e^x-e^{-x}}{e^x+e^{-x}}$，$relu(x) = max(0,x)$，$\sigma(x) = \frac{1}{1+e^{-x}}$，求得的表达式可复用$a,b,c \cdots y$简化）



（4）设（3）计算图中，输入$a=0,\ b=4,\ c=1$，参数$f=-3$，$y=1$时，计算$\frac{\partial L}{\partial k}$，$\frac{\partial L}{\partial c}$，$\frac{\partial L}{\partial a}$的值，如果$b=2$，$\frac{\partial L}{\partial c}$ 为多少？

二、 序列标注（25%）

(1)用 Viterbi 算法对以下分词后（用|隔开）的句子进行命名实体标注。假设标签集为{B-LOC, I-LOC, B-ORG, I-ORG, O}，并给出Viterbi算法的时间复杂度（假设标签集大小为m,序列长度为n），并指出抽出了几个实体，分别是什么。（提示：需给出最终标注结果及中间计算过程）

复旦 | 大学 | 位于 | 上海市 | 杨浦区 | 邯郸路 | 220 号

表 1 标签转移概率(例: P(B-LOC -> I-LOC) = 0.5)

标签	B-LOC	I-LOC	B-ORG	I-ORG	O
B-LOC	0.3	0.5	0.0	0.0	0.2
I-LOC	0.05	0.2	0.05	0.0	0.7
B-ORG	0.0	0.0	0.0	0.7	0.3
I-ORG	0.0	0.0	0.1	0.2	0.7
O	0.3	0.0	0.2	0.0	0.5

表 2 标签输出概率

	复旦	大学	位于	上海市	杨浦区	邯郸路	220号
B-LOC	0.0	0.0	0.1	0.9	0.8	0.8	0.3
I-LOC	0.0	0.0	0.0	0.0	0.1	0.0	0.7
B-ORG	0.8	0.3	0.0	0.0	0.0	0.0	0.0
I-ORG	0.0	0.7	0.0	0.0	0.0	0.0	0.0
O	0.2	0.0	0.9	0.1	0.1	0.2	0.0

(2)请根据自己的理解简述序列标注问题中为什么加入 CRF 能提升模型效果。

通过在大规模无标注语料上进行自监督学习，然后在少量下游任务数据上进行微调即可取得较高准确率。请你回答以下问题：

(1)请简述 ELMO,GPT,BERT 三者的异同点（可从模型架构、预训练任务等各种角度阐述），并从词向量的角度分析它们和 Word2vec,glove 这类方法有什么区别。

(2)最近以预训练模型为基础的 prompt 方法在小样本下游任务上取得了卓越效果，请简述什么是 prompt tuning,与常规的预训练模型 finetune 有什么区别，为什么 prompt 更适合小样本学习。

(3) 近期，美国人工智能公司 OpenAI 发布免费机器人对话模型 ChatGPT，模型目前处于测试阶段，用户与 ChatGPT 之间的对话互动包括普通聊天、信息咨询、撰写诗词作文、修改代码等，甚至有人觉得它未来可能取代 Google 成为新一代的搜索引擎。目前，其面世几天便已有超过 100 万用户使用，你是否尝试使用过并对其有所了解，请自由发挥谈谈你对 ChatGPT 的认识。（该小问为 bonus，共 5 分）

四、内容风控（25%）

知乎是国内流行的网络问答社区，其问题包含社会、时政、教育、情感等各个方面。如今，

知乎已经构建起坚实的内容壁垒。截至 2020 年 12 月，知乎上的总问题数超过 4400 万条，总回答数超过 2.4 亿条。据统计，知乎中 18-35 岁的青年占比约 75%，他们正处在人生的“第二个十八年”，正处在从认识世界迈向影响世界、改变世界的关键阶段，知乎的内容对他们的影响是巨大的，知乎的内容，决定着知乎的价值也引导着年轻人的价值观。因此，对知乎中的内容进行审核监管十分重要。过去，互联网公司通过增加内容审核人员规模来解决问题，比如 2018 年今日头条就曾将原有 6000 人的运营审核队伍扩大到 10000 人，社交巨头 Facebook 在全球范围内也拥有 1.5 万内容审核员。然而，人工审核本身是有缺陷的，比如说：审核成本高、审核慢、主观性较强，评价标准很难统一等。于是近几年，基于算法技术和人工智能，互联网平台开始开发机器辅助人工审核的方式应对内容安全问题。请运用你学过的自然语言处理知识回答以下问题：

(1) 请你从自己使用知乎等用户生成社区的经验出发，谈谈需要对用户生成内容进行何种审核，并设计相应方案实现对知乎平台发布的问题、评论进行自动内容审核(可综合多种方法)。

(2) 在系统上线后，可以利用用户与系统的交互记录来自适应地改进系统性能，比如可以记录某一问题的被点击次数来提升问题的推荐效果。请你在知乎平台设计一种方案使得内容风控系统能够基于线上用户行为实现持续学习。