

CSDN新首页上线啦，邀请你来立即体验！(http://blog.csdn.net/)

CSDN

博客 (/blog.csdn.net?ref=toolbar) 学院 (/edu.csdn.net?ref=toolbar)

下载 (/download.csdn.net?ref=toolbar) GitChat (/gitbook.cn/?ref=csdn)

更多 ▾



登录 (https://passport.csdn.net/account/login?ref=toolbar) 注册 (https://passport.csdn.net/account/mobileregister?ref=toolbar)

Attention Is All You Need

翻译 2017年11月25日 14:55:04

标签：机器翻译 (http://so.csdn.net/so/search/s.do?q=机器翻译&t=blog) /

attention (http://so.csdn.net/so/search/s.do?q=attention&t=blog) /

自然语言处理 (http://so.csdn.net/so/search/s.do?q=自然语言处理&t=blog) /

nlp (http://so.csdn.net/so/search/s.do?q=nlp&t=blog)

39

<https://arxiv.org/pdf/1706.03762.pdf> (https://arxiv.org/pdf/1706.03762.pdf)

摘要

主流的基于Encoder-Decoder的序列转换模型主要是基于复杂的递归或者卷积网络。现在好的模型还会加上一层聚焦(attention)机制。这篇文章我们提出一种新的网络框架，成为：Transformer，主要是基于attention机制，mn和cnn作为补充。这种方法在准确率和训练速度上面取得了相当不错的效果

介绍

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

立即体

验



显卡回收



人工智能学习



三星s6以旧换新



国外网站设计



广告 wendaJ (http://blog.csdn...)

+ 关注

(http://blog.csdn.net/chazhongxinbitc)

码云

0

原创

粉丝

喜欢

6

1

0

(https://git

utm_sour

他的最新文章

更多文章 (http://blog.csdn.net/chazhongxinbitc)

QA (三)：复杂attention机制(coattention及bi-attention) (http://blog.csdn.net/chazhongxinbitc/article/details/78825704)

瀑布流排序中的position偏置消除的实验 (http://blog.csdn.net/chazhongxinbitc/article/details/78812090)

QA(二)：利用Attention机制，带着问题阅读 (http://blog.csdn.net/chazhongxinbitc/article/details/78724911)

Image captioning(三) (http://blog.csdn.net/chazhongxinbitc/article/details/78689754) 登录 注册



内容举报



返回顶部



Rnn,特别是LSTM,GRU这种gate机制的网络在过去一段时间被证明在sequence model 或者 transduction 问题上取得了非常不错效果,他是 Encoder-Decoder 的基础

rn的机制,对sequence里面的每一个元素递归迭代得到一个状态 $h(t)$,然后 $h(t)$ 是下一乱迭代的一个输入,那它天然无法并行,所以使用rn训练会更慢,时间更久,特别是当sequence很长的时候。最近人们用因子分解的形式来加快训练速度,但是性能瓶颈依然存在。

Attention 现在已经集成在sequence model 或者是 transduction问题中。它的作用用户解决input sequence 和 output sequence的位置不是一一对应的,甚至他们的长度,顺序等都是不一致的问题,Attention可以捕获位置信息。然后这中形式的Attention是伴随着RNN的计算生成的(如果用卷积encoder的话就是伴随卷积层生成的),主要它不是独立的。

在我们当前的模型Transformer,我们去递归,相应的我们全部依赖Attention。主要是我们要在input和output之间建立一个全局的依赖。这种机制兼顾了并行和效果。

背景

类似降低序列计算量的基础工作包括:Extended Neural GPU,ByteNet,ConvS2S,都是基于卷积的方式,在这些模型中,他们是要学习input和output对等元素的相对位置表示,那么如果两种场景的元素距离很遥远,那么计算量就会随之线性上升(通常要对每个相对位置计算softmax,所以计算量很大)。在这个模型中,我们把计算量控制在常量范围,我们这么做是有效率下降的,因为我们把 attention-weight做了平均,比如:Multi-Head Attention

Self-attention,它用在单个sequence时,能够捕获位置信息,这样学习出来的表示更加准确,它在阅读理解,文本摘要,句子表示中取得了不错的成绩。

前面说的这些attention机制,都是基于其他的encoder技术,Transformer是第一完全依靠Self-attention机制来表示

模型

加入CSDN,享受更精准的内容推荐,与5000万程序员共同成长!

Image captioning(二) (<https://passport.csdn.net/account/mobile>)
net/chazhongxinbitc/article/details/78631849



显卡回收



人工智能学习

Attention is
试 (<http://blog.csdn.net/lqfarmer/article/details/73521811>)



三星s6以旧换新



国外网站设计

模型汇总16 各类Seq2Seq模型对比及《Attention Is All You Need》中技术详解 (<http://blog.csdn.net/lqfarmer/article/details/73521811>)

谷歌机器翻译Attention is All You Need (<http://blog.csdn.net/dellme99/article/details/74066975>)

一三五浪可加长, 每段细分五小浪
另有等长九段波, 顶底不连通连长
三三相隔十五段, 五三交错亦寻常
波起浪伏有形状, 常见上斜与扩张
喇叭斜三现一浪, 二浪之后走势强
五浪若是此模样, 分批减持远危墙
A 浪止住回头看, A 3 A 5 不一样
三波之字双回撤, 五波右肩做 B 浪
回撤二次分三五, 三波弱来五波强
B 浪右肩 a-b-c, 趁金快手捕长阳

散户炒股口诀

全透明手机

学唱歌先练什么

loft 公寓

内容举报

返回顶部

登录

注册

X

现在大部分的转移模型都是基于 encoder-decoder,encoder过程是将 sequence的序列表示 (x_1, x_2, \dots, x_n) 转换成一个连续的序列

表示 $z = (z_1, z_2, \dots, z_n)$, decoder过程是更具 z 生成对应的

output序列 y_1, y_2, \dots, y_m .Transformer参照这个整体的框架,

主要处理时用到了 Self-attention 和 point-wise , encoder 层和 decoder层是使用全链接

encoder 和 decoder 层

Encoder: 由6层组成, 每个层有两个子层sub-layer, 第一个子层是 multi-head self-attention , 第二个子层是 point-wise前向全链接层。每个子层之间我们用残差链接, LayerNorm(x+ sub-layer(x)), LayerNorm 为正则化。 encoder的整体输出维度为 512

Decoder: 和 encoder 大致类似, 每层有三个子层 sub-layer构成, 前两层一样, 新增加的一层multi-head self-attention是作用于Encoder的输出, 有一点不一样的是: output连接的multi-head self-attention层要注意又个 masking操作, 他的意思是我预测position i, 只能用 i-1 之前的位置信息, 这个很容易理解, predict的过程中, 我们是按sequence来生成的, 预测 word i的时候, 我们只能用之前的信息, 不能跨越, 用后面的信息。

Attention

首先简单介绍下之前其他的encoder-decoder 模型的 Attention机制:

简单的Attention作用在decoder阶段, 它主要解决的问题是: 假设给定一个原始句子"what are you doing" encoder 阶段会会有两个数据产出: 经过每个词后的向量产出: attention_outputs(向量数量等同与词的个数), 和处理完所有词之后保留的一个固定的状态信息向量: encoder_state

decoder阶段的输入分为三个: 第一前一个预测出来的词是什么, 如果是第一个词, 那么用一个固定的初始词,

第二是 encoder_state+经过前n个词生成的状态信息encoder_state

第三是 attention_outputs

加入CSDN, 享受更精准的内容推荐, 与5000万程序员共同成长!



他的热门文章

QA(二): 利用Attention机制, 带着问题阅读 (<http://blog.csdn.net/chazhongxinbitc/article/details/78724911>)

63

Image captioning (一) (<http://blog.csdn.net/chazhongxinbitc/article/details/78689456>)

49

Dynamic Routing Between Capsules (<http://blog.csdn.net/chazhongxinbitc/article/details/78631354>)

43

QA: Dynamic Memory Networks for Natural Language Processing (<http://blog.csdn.net/chazhongxinbitc/article/details/78686730>)

42

Attention Is All You Need (<http://blog.csdn.net/chazhongxinbitc/article/details/78631849>)

37

内容举报

返回顶部

登录

注册

X

假设decoder阶段已经翻译出来：“你 在 做”，然后预测下一个词“什么”

这个工作输入为：

encoder_state(encoder_state 在经过：你 在 做的处理之后保留的状态)

attention_outputs，这里的原始信息是 state的状态和 英文原始的信息共同决定 下一个词最可能是什么，state存储的是整体内存信息，attention_outputs是决定当前的词和原始句子中的那个词最相关。

这里的Attention要复杂一点，主要是对encoder 和 decoder阶段的 词进行更加复杂化的处理，这种复杂化甚至可以取代原来基于rnn或者cnn 的encoder 和 decoder操作

其实我觉的只是玩了一个概念上的东西而已，并没有改变 e-d 这种框架

第一步：encoder阶段生成一个固定维度的向量

第二部：decoder阶段利用encdoer阶段生成的向量，然后结合自己的一个组织方式，预测序列

Scaled Dot-Product Attention

这个是本文用到的 Multi-Head Attention 中的一步：

Scaled 的意思是按比例增加或者缩放向量，比例为： $\sqrt{d_k}$,

d_k 是 word embedding 的维度

这里引入 scale的作用是避免维度过大带来两个向量dot时的数据过大，从而被clip掉

这里为什么 $\sqrt{d_k}$ ，是因为如果每一个值服从正态分布，他们的整体标准差是 $\sqrt{d_k}$

Dot-Product：我理解为各种复杂的 点击 矩阵相乘计算

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

(<https://passport.csdn.net/account/mobile>)



显卡回收



人工智能学习



三星s6以旧换新



国外网站设计

广告



0



内容举报



返回顶部

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

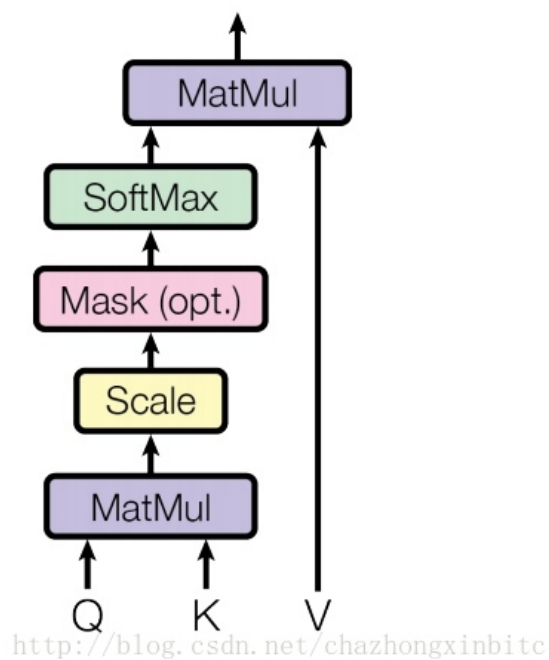
登录

注册





Scaled Dot-Product Attention



解释下这个图：

整体的是：

self的体现：k, v 一般是一样，或者相关的矩阵，利用query 和 k 作用后(比如矩阵相乘，softmax归一等) 出来的值k1(类似 rnn 生成状态向量 state)，然后和 k1 和 v(类似rnn的attention_outputs) 相乘后 既保留state信息，又保留原始 短语机构信息，

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

(<https://passport.csdn.net/account/mobile>)



显卡回收



人工智能学习



三星s6以旧换新



国外网站设计

广告



内容举报



返回顶部

登录

注册



输入：Q, k, v 是经过计算后的 query, 和 k, v 矩阵

MatMul: QK^T query 和 k 做矩阵相乘

scale: $\frac{QK^T}{\sqrt{d_k}}$ 每个元素除以 $\sqrt{d_k}$

Mask: 因为计算是基于矩阵的方式，所以batch 操作的时候有个补足对齐的工作，为了避免补足这样的词的位置影响到predict，所以我们计算的时候只考虑seq的原始大小，对于不足的数据，我们直接用 负无穷来重置，负无穷经过 softmax之后基本就是0，不会对后续产生影响

SoftMax: 这里是基于最低的维度做softmax，保留输入矩阵的结构，最快就是简单的归一化

MatMul: 左边 $\text{softmax}(\frac{QK^T}{\sqrt{d_k}})$, 右边 V

整体思想：

1. 假设Q 和 K, V都是一个矩阵，

权重计算C：Q 和K 矩阵相乘后是一个权重矩阵，假设Q 和K的矩阵都是归一化的，那么Q中的第i行向量和K中的第i列向量是相同的，他们做点积后是这行里面最大的，其他的元素可以行量第i个词和其他词之间的距离和相似性，这里就体现了self的概念，先学习到自身词之后的联系，后续softmax 起到归一化的作用

计算影响的状态：CV, C是圈中，V是内容自身，权重和自身内容作用，得到内容里面重要的信息

这里就相当于卷积做pooling的变形版本 $A * V$, A如果是常量就是fastText，A如果是一个train的变量那么就是简单的卷积Encoder，

这里 $A = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})$ ，然后叫 self attention

1. 假设Q 和(K, V)不是一个矩阵，Q是原始序列矩阵，K, V是翻译后的序列矩阵，一般而言 Q是decoder的输入， K_1, V_1 是

权重计算 C_1 ：这里有个假设，不同语言中的同一词(比如中文：男人，英文：man)在它本身词库中的向量位置是差不多的，那基于这个假设，Q 中的1个向量在和 K 中的每个向量做点积的时候，结果数值最高的向量应该就是其他语言中和这个词对应的词，也可以得到其他的词和这个词之后的距离关系，然后经过softmax，这样的权重在不同的语言之间也可以计算了

计算影响的状态： $C_1 V_1$

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

(<https://passport.csdn.net/account/mobile>)



显卡回收



人工智能学习



三星s6以旧换新



国外网站设计

广告



内容举报



返回顶部

登录

注册



这里既然是翻译：那么两种言语对应的内容应该是一致的 假设

那么由1可以计算得到内容：

对于en的seq的自身内容提取： $context = CV$

对于de的seq的自身内容提取： $context_1 = C_1 V_1$

根据2的理论：

如果信息完全 $context$ 和 $context_1$ 应该表达的信息对等，我们利用 $context_1$ 矩阵做softmax predict就可以预算

下一个词应该翻译成什么

但是翻译的时候 因为内容未知， $context_1 = C_1 V_1$ 的内容是不完整的，但是内容表达部分我们可以借助

$context$ 来表示，假设我们已经翻译出“你 在 做”，然后利用self attention 得到 $context_1$ ，那利用理论2，我

们可以结合 $context_1$ 和 $context$ 计算和当前两个词相关的 $context$ 内的部分，这样的到的状态向量用来做

softmax

Multi-Head Attention

Multi-Head是基于上面的实例，对每个矩阵可以进行分割，然后做完Dot-Product Attention后再合起来，多个分割矩阵总比多个Head的Attention表示多了很多可能性

在模型中应用Attention

(<https://passport.csdn.net/account/mobile>)



显卡回收



人工智能学习



三星s6以旧换新



国外网站设计

广告



0



内容举报



返回顶部

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

登录

注册



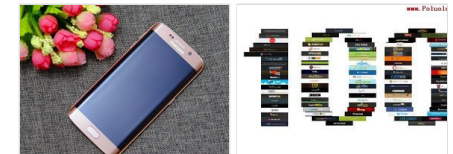


(<https://passport.csdn.net/account/mobile>)



显卡回收

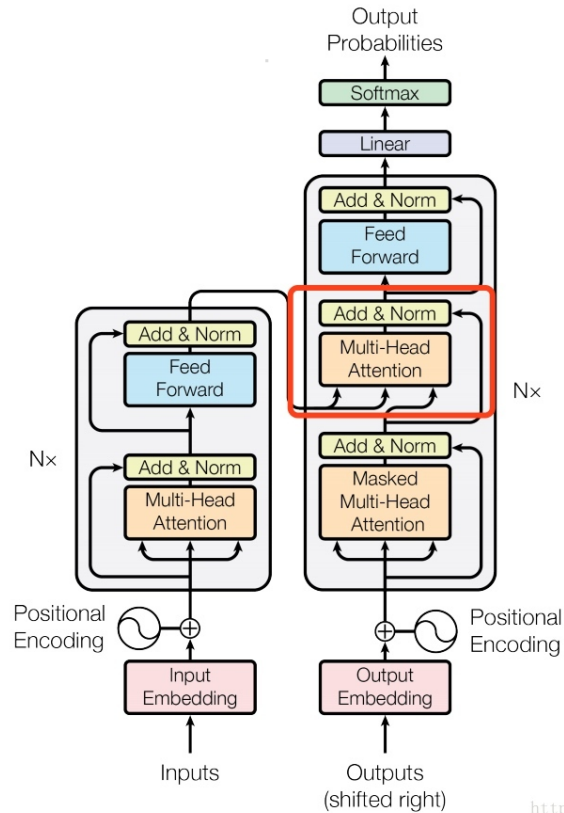
人工智能学习



三星s6以旧换新

国外网站设计

广告



<http://blog.csdn.net/chazhongxinbitc>

Transformer 在使用Attention有三个特别的地方

1. 在红框出的Attention中，query是来自上一个decoder层self Attention输出，key 和value是来自encoder层的最终输出，这样就可以是 decoder层的每一个position和 encoder层的所有位置建立计算关系。这是和seq2seq最大的不同的地方。
2. encoder 层有一个self Attention的机制，他们的输入query， key， value都是由上一个层输出的。这样每个encoder层的position都会和上一个encoder的所有位置信息建立联系。

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！



内容举报



返回顶部

登录

注册



3. 同样decoder层是一样的。为了避免自回归(比如无效的词或者矩阵对齐的词),就是通过softmax层的时候影响向量分布,所以在进入softmax之前将该值设置为负无穷,这样经过softmax之后,该值就为0了。这样就避免了建立一些非法链接

Position-wise Feed-Forward Networks

在encoder和decoder层的最终结果都会经过一个Feed-Forward层。这块主要是两个线性全链接层

输入: input。

第一个加了一个Relu激活函数 $\text{output1} = \max(0, \text{xW1} + \text{b1})$

第二个就是一个线性层 $\text{output} = \text{output1} * \text{W2} + \text{b2}$

最终输出: $\text{output} = \text{output} + \text{input}$

也可以用卷积的形式进行处理

Embeddings and Softmax

词的向量用已经训练过的向量

在embedding layer 可以每个元素乘以 $\sqrt{d_{\text{model}}}$, d_{model} 为词向量的维度

Positional Encoding

因为没有利用递归和卷积,为了学习词序这些信息,我们必须利用词的position信息,所以考虑将position惊醒encoding

为什么 self-Attention

主要用self-Attention 替代递归或者卷积队自身的信息进行表达,原因有3个:

1. 整体的计算复杂度
2. 实现并行化
3. 句子中有依赖的词之间距离可能很远,怎么去学习很远距离的两个词的关系

加入CSDN,享受更精准的内容推荐,与5000万程序员共同成长!

(<https://passport.csdn.net/account/mobile>)



显卡回收



人工智能学习



三星s6以旧换新



国外网站设计

广告



内容举报



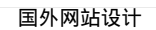
返回顶部

登录

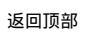
注册



(<https://passport.csdn.net/account/mobile>



X





相关文章推荐

Attention is all you need新翻译架构的测试 (http://blog.csdn.net/sparkexpert/article/detail...

翻译的进展真是很快，如近日，谷歌再次宣布又在机器翻译上更进了一步，实现了完全基于 attention 的 Transformer 机器翻译网络架构。这篇文章的模型完全是在编码 - 解码程序基础上加上A...



sparkexpert (http://blog.csdn.net/sparkexpert) 2017年06月27日 09:04 562

对Attention is all you need 的理解 (http://blog.csdn.net/mijiaoxiaosan/article/details/7325...

对谷歌Attention is all you need 的理解。



mijiaoxiaosan (http://blog.csdn.net/mijiaoxiaosan) 2017年06月14日 19:24 5940



太任性！学AI的应届学弟怒拒20K Offer，他想要多少钱？

AI改变命运呀！！前段时间在我司联合举办的校招聘会上，一名刚刚毕业的学弟陆续拒绝2份Offer，企业给出18K、23K高薪，学弟拒绝后直接来了一句...

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjnvPjn0IZ0qnfk9ujYzP1f4PjDs0Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1YLuymLujDsPhw9PjN-rAR30AwY5HDdnHfsnWfdnHD0lgF_5y9YIz0IQzq-uZR8mLPbUB48ugfEIAqspynEmybz5LNYUNq1ULNzmvRqmhkEu1Ds0ZFB5HD0mhYqn0KsTWYs0ZNGujYkPHTYn1mk0AqGujYknWb3rjDY0APGujYLnWm4n1c0ULI85H00TZbqnW0)

模型汇总16 各类Seq2Seq模型对比及《Attention Is All You Need》中技术详解 (http://blog.c...

1、已有Seq2Seq模型 Seq2Seq模型是处理序列到序列问题的利器，尤其是在神经网络翻译（NMT）方面，取得了很大的成功。Seq2Seq由一个encoder和一个decoder构成，encode...



lqfarmer (http://blog.csdn.net/lqfarmer) 2017年06月20日 22:00 1525

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

(https://passport.csdn.net/account/mobile



显卡回收



人工智能学习



三星s6以旧换新



国外网站设计

广告

内容举报

TOP

返回顶部



登录

注册

X

谷歌机器翻译Attention is All You Need (<http://blog.csdn.net/dellme99/article/details/74066...>)

https://mp.weixin.qq.com/s?__biz=MzI3MTA0MTk1MA==&mid=2651999438&idx=1&sn=9ede16ab3de8ded3b603870ba9...

 dellme99 (<http://blog.csdn.net/dellme99>) 2017年07月01日 16:29  758

(<https://passport.csdn.net/account/mobile>)



显卡回收

人工智能学习



三星s6以旧换新

国外网站设计

广告

All You Need To Know About Windows Phone 8 (<http://download.csdn.n...>)





<http://download.csdn.net/detail/dellme99/74066...> 2013年09月16日 11:02 11.32MB [下载](#)

 <p>1 0.25/个 批量改价.USB电线扣. 汽车线卡.汽车电线固</p>	 <p>2 45.00/米 乐品-塑料光纤槽 道-120*100mm</p>	 <p>3 288.00/件 KSS线槽VD10实际库 存促销, 原装正品欲购</p>
---	---	--



Request caching is not available. Maybe you need to initialize the HystrixRequestContext...

在《spring cloud 微服务实战》书中第159页-----请求缓存这一部分，通过继承HystrixCommand的方式实现的命令，开启请求缓存只需通过重载getCacheKey()方法， @...

 lvyuan1234 (<http://blog.csdn.net/lvyuan1234>) 2017年08月04日 18:48  823



Life is short, You need Python (<http://blog.csdn.net/u011012932/article/details/52486082>)

『人生苦短，我用 Python』，作为一个 Pythoner，这句话再熟悉不过了。一起用心来感受下吧！只看图，不说话。 ...

 u011012932 (<http://blog.csdn.net/u011012932>) 2016年09月09日 13:00  4738

【原】The 'InnoDB' feature is disabled; you need MySQL built with 'InnoDB' to have it w...

今天安装php程序的时候，突然mysql报出了个错误：The 'InnoDB' feature is disabled; you need MySQL built with 'InnoDB' to h...

 xiaobing_122613 (http://blog.csdn.net/xiaobing_122613) 2017年01月23日 14:12  136

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长。
Even All I See Is You 翻译 (<http://download.csdn.net/download/mmmm...>)



内容举报




返回顶部

登录


注册






hadoop-what-you-need-to-know.pdf (http://download.csdn.net/downloa...)

2017年12月14日 20:16 9.95MB 下载




This application everything you need.t has a CALCULATOR.It h (http://d...)

2006年02月23日 09:05 29KB 下载



10.Lessons.About.C++.You.Need.To.Learn.To.Become.A.Master.Progra...

2015年03月25日 14:10 5.57MB 下载



You Probably Don't Need RAC (http://download.csdn.net/download/kele...)

2010年04月16日 12:13 117KB 下载

failed to sync branch You might need to open a shell and debug the state of this repo. (h...

github同步失败

dongqinliuzi (http://blog.csdn.net/dongqinliuzi) 2015年07月04日 20:13 5106

URAL 1993-This cheeseburger you don't need (模拟) (http://blog.csdn.net/u013534690/...

Description Yoda: May the Force be with you. Master Yoda is the oldest member of the ...

u013534690 (http://blog.csdn.net/u013534690) 2014年07月28日 22:22 473

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

(https://passport.csdn.net/account/mobile



显卡回收



人工智能学习




三星s6以旧换新




国外网站设计



广告



内容举报




返回顶部

登录 注册

×

QT 5.7 for iOS Xcode 8 Project ERROR: Xcode not set up properly. You may need to con...

手机升级到了ios10，然后想着懒得折腾直接升级到xcode 8好直接真机调试，嗯，想法是对的，然后xcode 8上也可以直接在ios 10上调试了。但是当换到Qt creator 4.0.1 / ...

 Enter_ (http://blog.csdn.net/Enter_) 2016年10月16日 10:49 3342

(<https://passport.csdn.net/account/mobile>)



显卡回收



人工智能学习



三星s6以旧换新



国外网站设计

广告



0



[Vista基础教程] 100 Things You Need to Know about Microsoft Windows...

<http://download.csdn.net/detail/20080420/2232000> 2008年04月20日 09:59 22.32MB [下载](#)



What you need to know about Angular 2 (<http://download.csdn.net/detail/20161110/626000>)

<http://download.csdn.net/detail/20161110/626000> 2016年11月10日 09:53 626KB [下载](#)

you need to use a theme.appcompat theme (or descendant) with this activity 解决办法 (h...

当你想隐藏 Androidmanifest.xml android:n

 xuqingfeng77 (<http://blog.csdn.net/xuqingfeng77>) 2015年05月28日 11:01 7649



内容举报



返回顶部

加入CSDN，享受更精准的内容推荐，与5000万程序员共同成长！

登录

注册

