深度之眼
deepshare.net

本期论文主题:Transformer

导师：Yamada

《Attention is all you need》

注意力机制是大家需要掌握的

作者：Ashish Vaswani

单位：google

发表会议及时间：NIPS，2017

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

RNN/LSTM

了解循环神经网络（RNN/LSTM）的结构，掌握RNN的基本工作原理

Seq2Seq

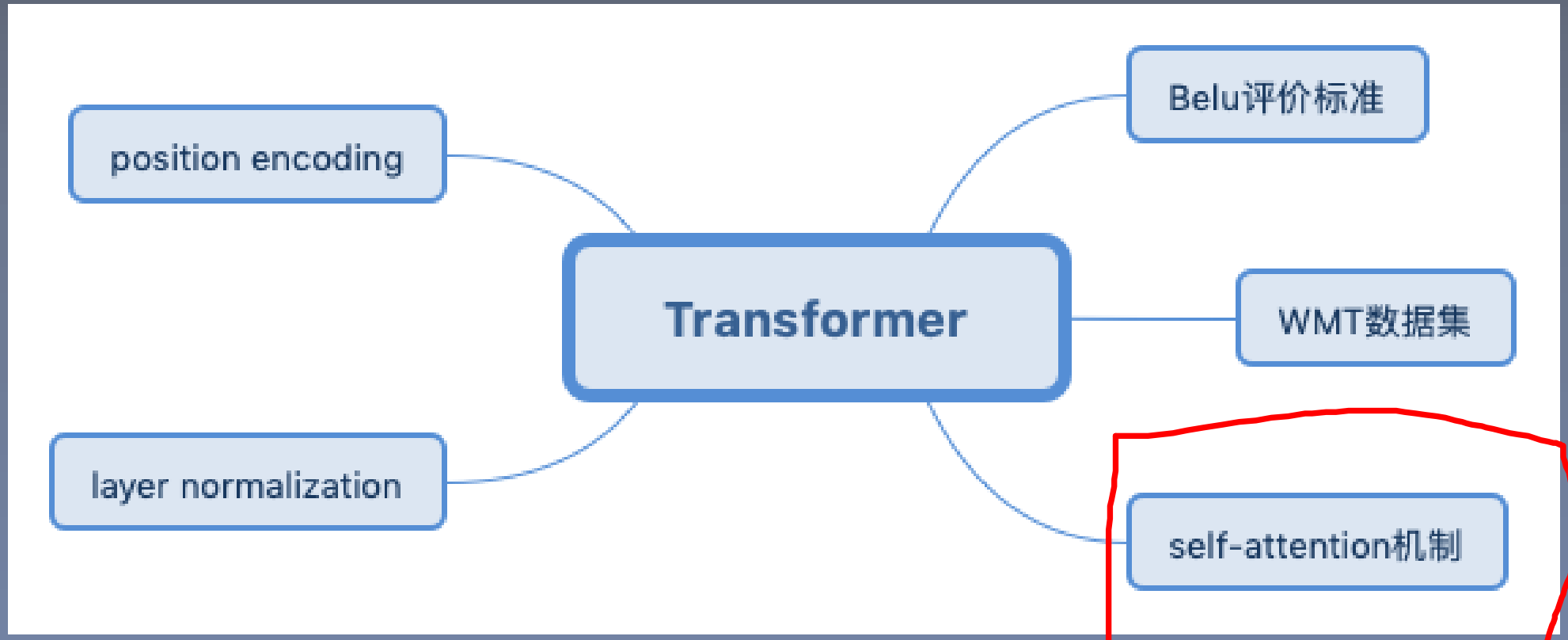
了解Seq2Seq的概念，掌握Seq2Seq的基本工作原理

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
论文总览
Transformer结构

03

Part 3
self-
attention

04

Part 4
小trick

05

Part 5
论文总结

06

Part 6
代码实现分析

深度之眼
deepshare.net

第一课：论文导读

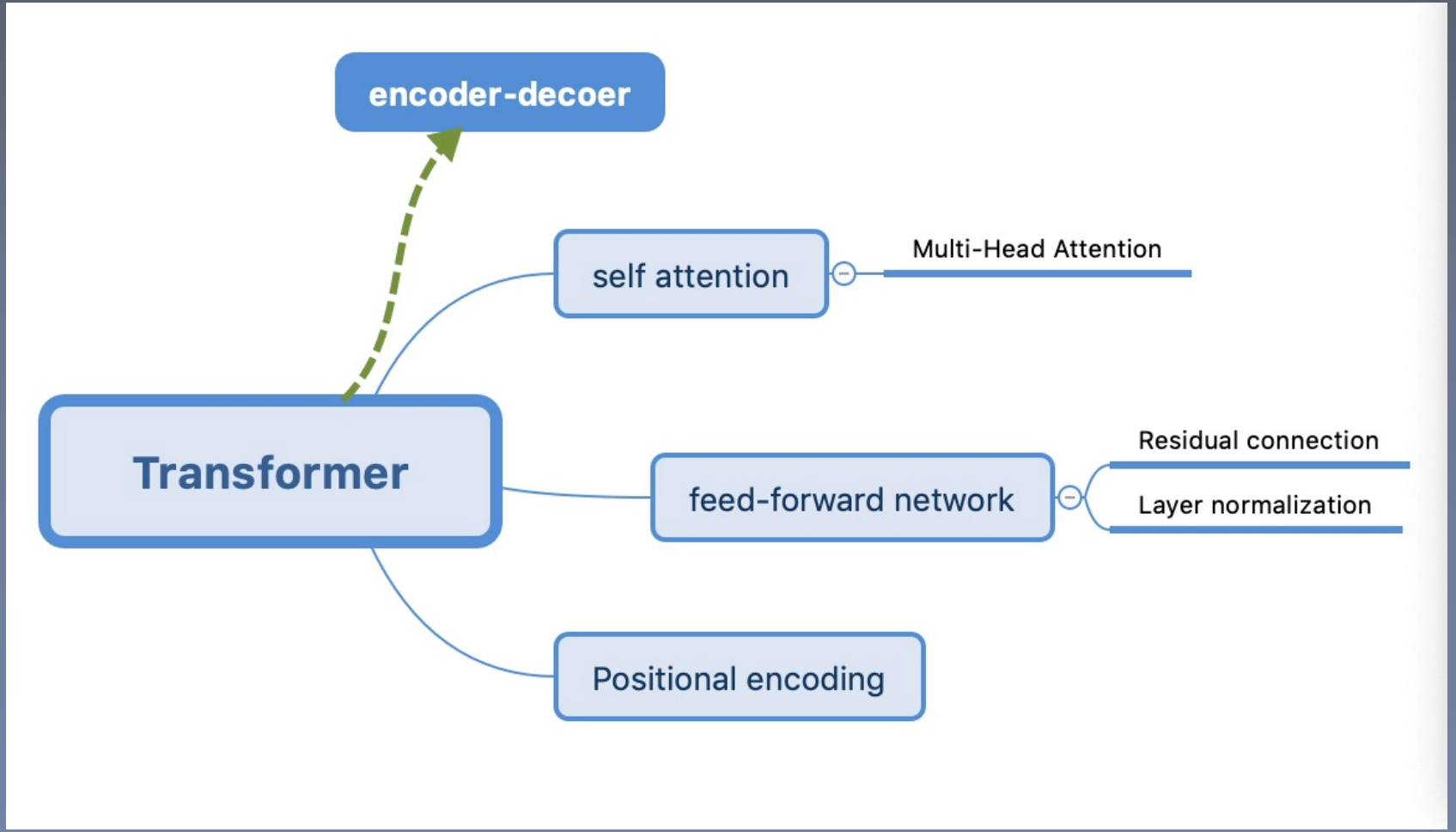
The first lesson: the paper guide

深度之眼
deepshare.net

目 录

知识树

深度之眼
deepshare.net

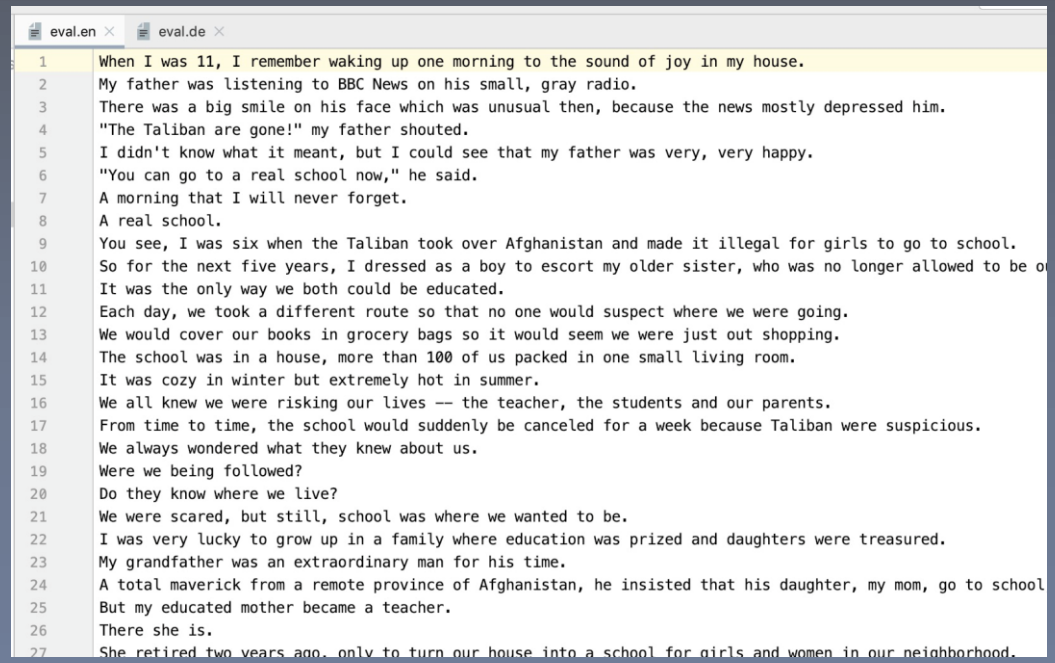


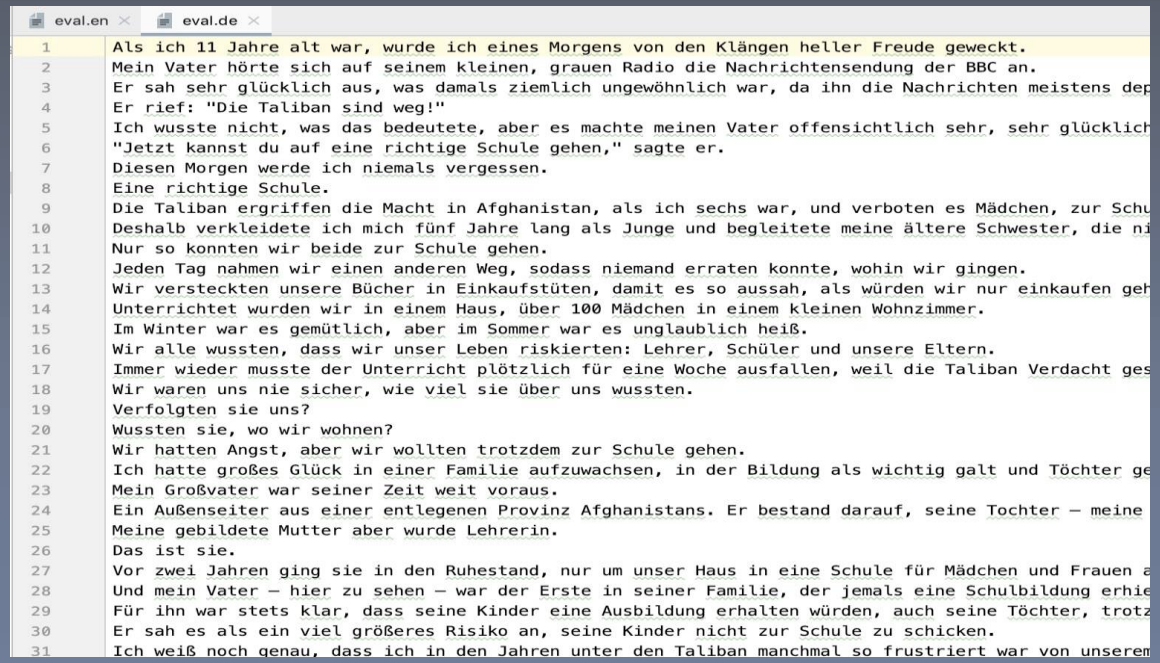
深度之眼
deepshare.net

论文研究背景、成果及意义

研究背景

Research background





WMT翻译数据集

wmt数据集包括德语翻译成英语、法语翻译成英语等数据集。数据集级量级在百万级别。

研究背景

Research background

深度之眼
deepshare.net

翻译效果衡量指标bleu

重点 重点来了！

bleu采用了一种N-gram的匹配规则，去比较议文和参考译文n组词的相似比。

参考译文 It is a nice day today

译文 Today is a nice day

5/6

2/4

研究背景

Research background

深度之眼
deepshare.net

翻译效果衡量指标bleu

重点 重点来了！

参考译文：the the the the

译文：The cat is standing on the ground

$$Count_{clip} = \min(Count, Max\_Ref\_Count)$$

→ 2

$$P_n = \frac{\sum_i \sum_k \min(h_k(c_i), max_{j \in m} h_k(s_{ij}))}{\sum_i \sum_k \min(h_k(c_i))}$$

$$BLEU = BP \times exp(\sum_{n=1}^{N} W_n log P_n)$$

研究成果

Research Results

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

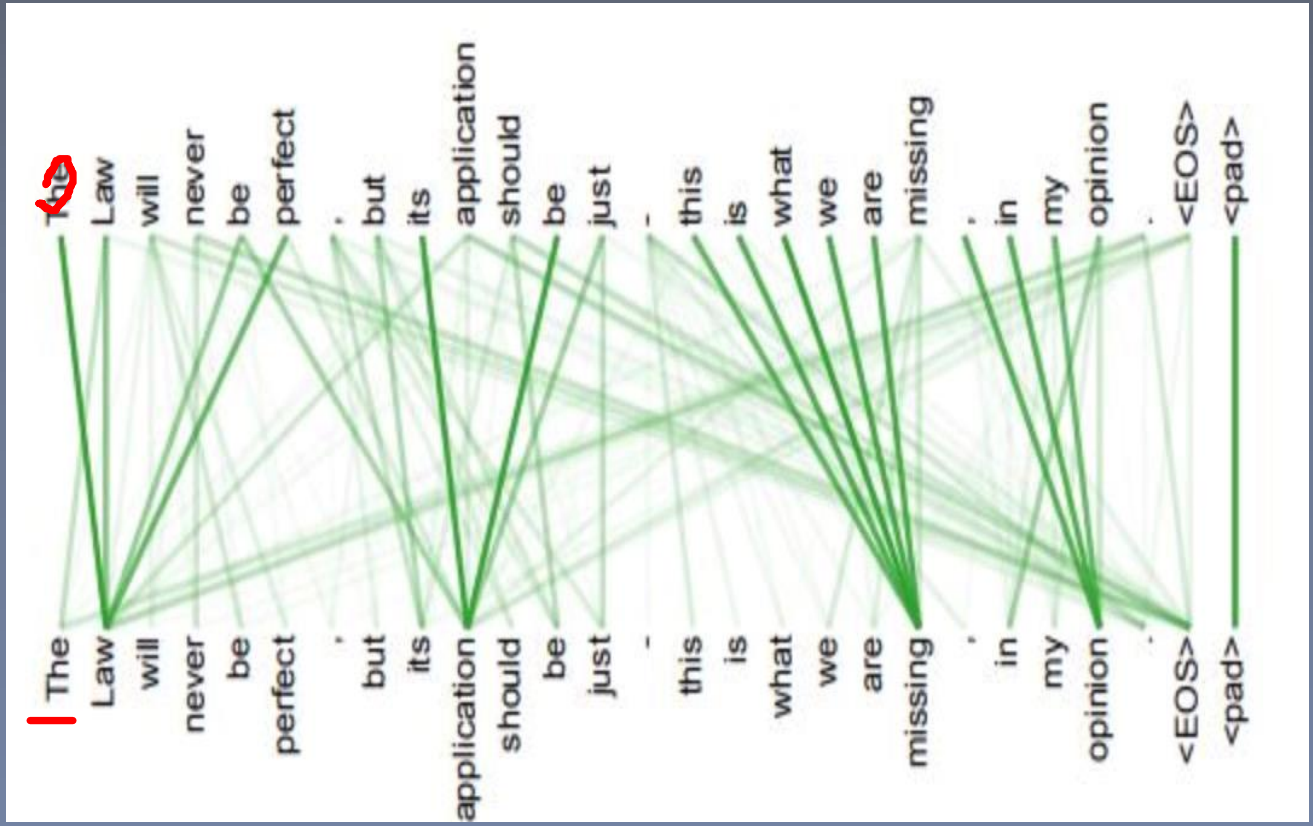
Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

1 在WMT 2014English-to-German 翻译任务上比其它模型的bleu值高出两个点。

2 时间复杂度上和传统模型相比大大降低，还可以用于并行。

研究成果

Research Results



3 self-attention模型具有更强的可解释性，左图attention结果显示了不同词语之间的关联信息。

研究意义

Research Meaning

重点 重点来了！

Transformer历史意义

- **提出self-attention，拉开了非序列化模型的序幕。**
- **为预训练模型的到来打下了坚实的基础**

nlp领域

序列化模型占主导地位

LSTM为代表

2017

里程碑Transformer

nlp领域

提出一种新的attention方式，实现非序列化模型的并行化，提高效率

self-attention

研究意义

Research Meaning

Transformer历史意义

- **提出self-attention，拉开了非序列化模型的序幕。**
- **为预训练模型的到来打下了坚实的基础**

基于transformer结构的预训练模型：

bert(采用的transformer的encoder部分)

gpt(采用的transformer的decoder部分)

albert等tingybert模型。

深度之眼
deepshare.net

论文泛读

Strcuture of Paper

论文结构

Structure of Papers

深度之眼
deepshare.net

摘要 Abstruct

介绍背景及提出Transformer模型，在WMT翻译数据集上比最好模型的bleu值高出两个百分点

1 Introduction

提出序列化模型的弊端；并提出本文的模型

2 The Backgroud

传统卷积模型存在学习远距离依赖困难的弊端，并提出自注意力机制

3 The Architecture

Transformer网络结构及其内部细节：Scaled Dot-Product Attention、Multi-Head Attention、Positional Encoding.

4 Why Self-attention

分析self-attention和cnn、lstm的时间复杂度。

5 Training

训练语料以及硬件、超参数设置介绍

6 Results

Transformer的实验结果对比

7 Discussion

计划将transformer结构应用于除翻译外的其它领域

摘要
abstract

摘要核心

1. 常用的序列模型都是基于卷积神经网络或者循环神经网络，表现最好的模型也是基于encoder-decoder框架的基础加上attention机制。

2. 提出一种基于attention机制的新模型transformer，抛弃了传统的模型结构。

3. 模型在2014WMT翻译数据集上，比现存最好的模型的bleu值高2个点。

论文小标题

Paper title

深度之眼
deepshare.net

深度之眼
deepshare.net

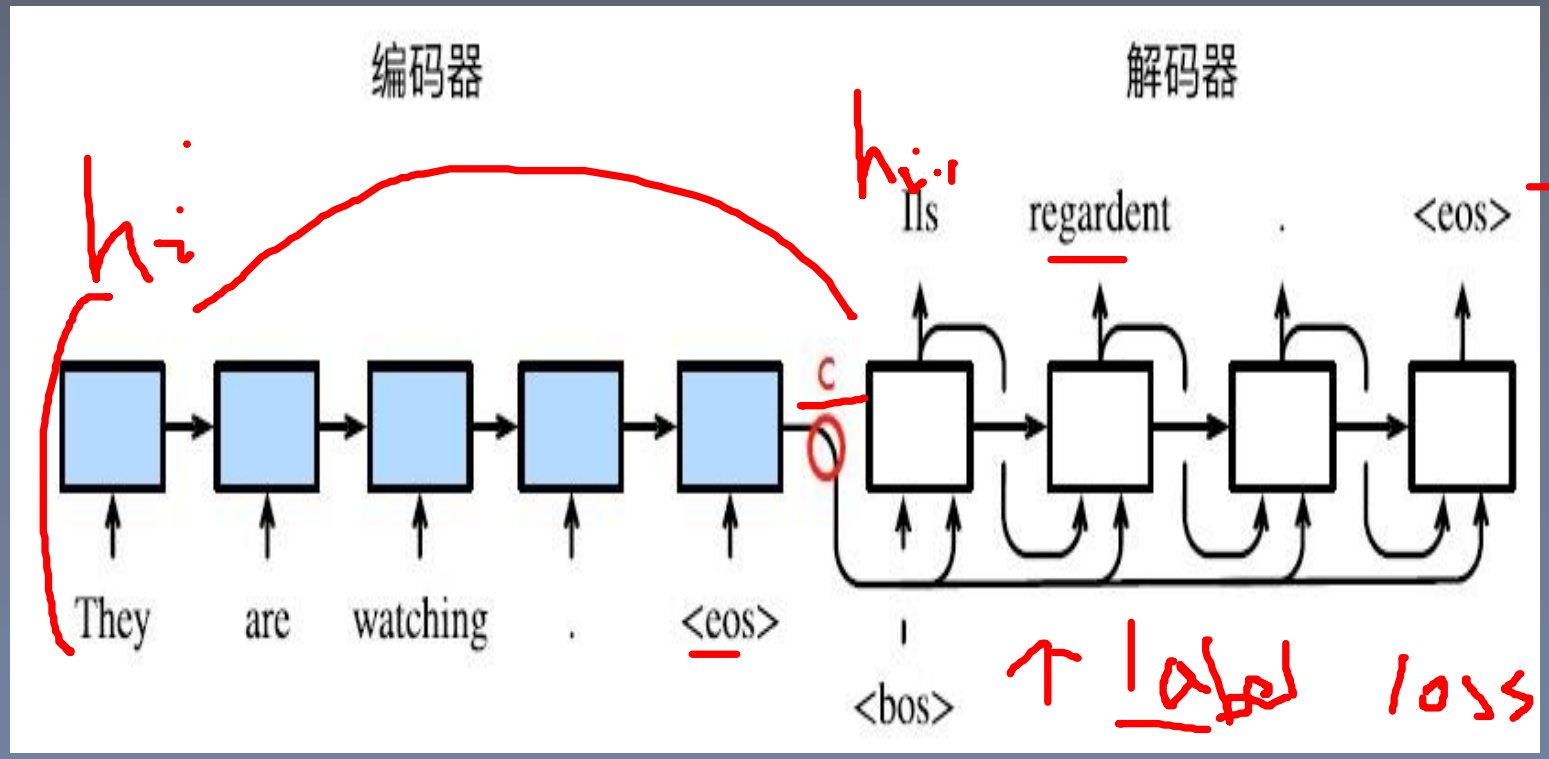
seq2seq以及attention的回顾

Strcuture of Paper

论文结构
Structure of Papers

深度之眼
deepshare.net

seq2seq模型以及attention机制



$$Attention(Q, K, V) = softmax(QK^T)V$$

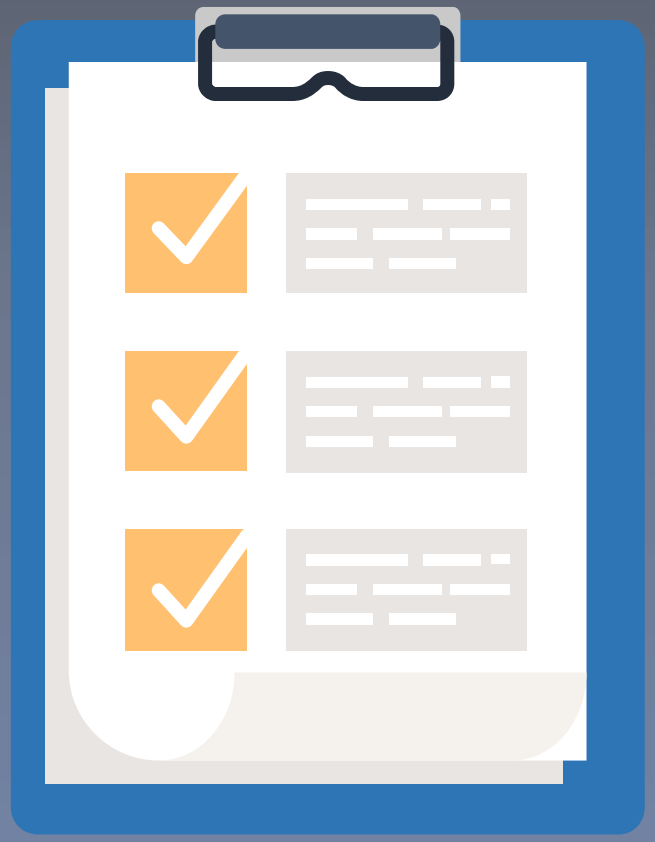
深度之眼
deepshare.net

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



01 课程安排

3个课时，导读、精读、代码，6个部分。

02 论文总览

论文总共包含7个部分，论文主要介绍self-attention机制并验证其有效性。

03 研究背景及成果意义

学习到一个机器翻译数据集的衡量指标bleu。
Transformer和预训练模型之间的关系。

深度之眼
deepshare.net

深度之眼
deepshare.net

下节预告

Preview of next lesson

01 Transformer结构

讲解Transformer的结构，为什么transformer可以解决传统序列化模型的获取长文依赖难？

02 self-attention结构

讲解self-attention的结构，self-attention和传统的attention机制有什么不一样的地方

03 实验设置及结果分析

网络超参数设置，学习率，batchsize等
实验结果分析对比

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

深度之眼
deepshare.net

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
Q Q： 2677693114

公众号

客服微信