

# 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



# 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

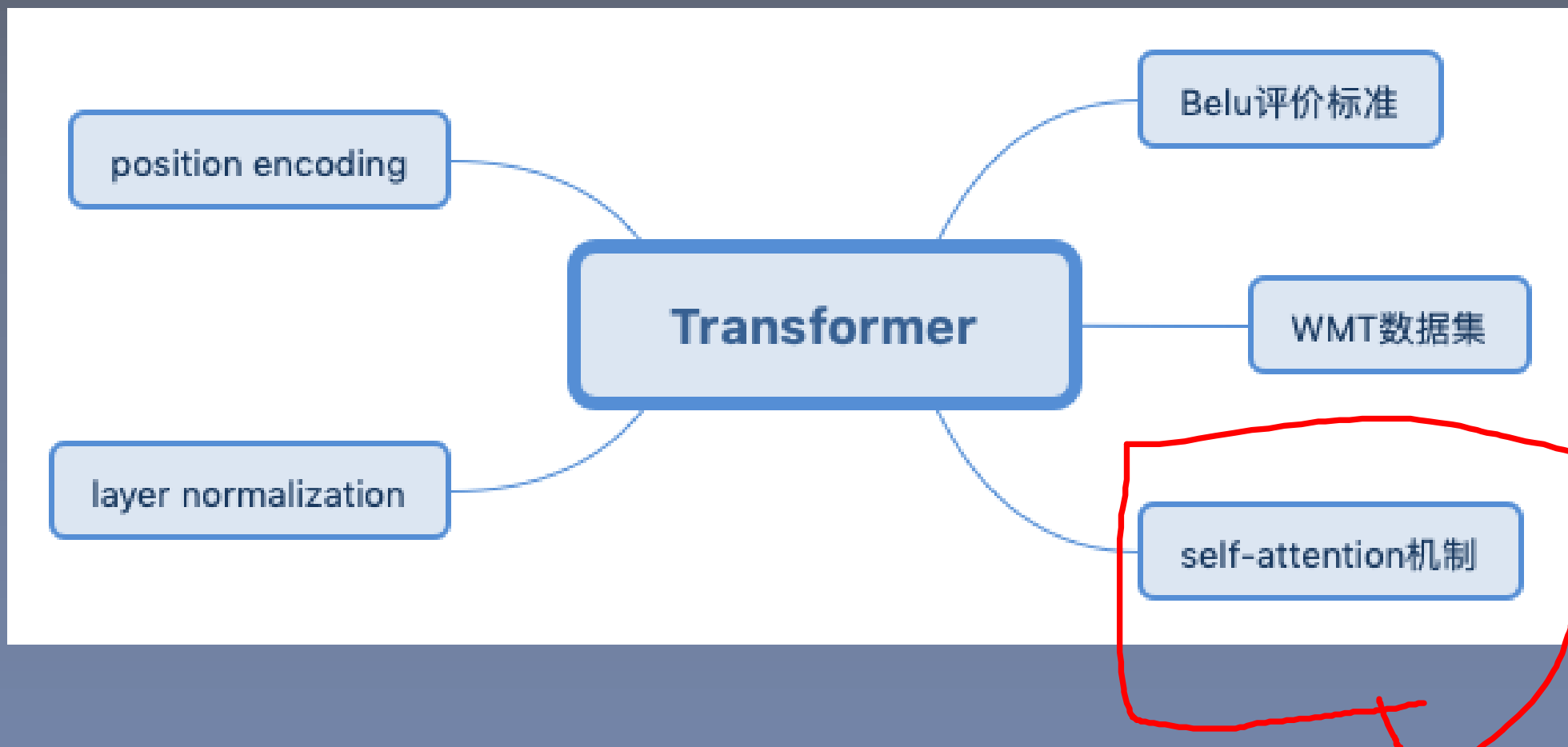
了解Seq2Seq的概念，  
掌握Seq2Seq的基本工作原理

## 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

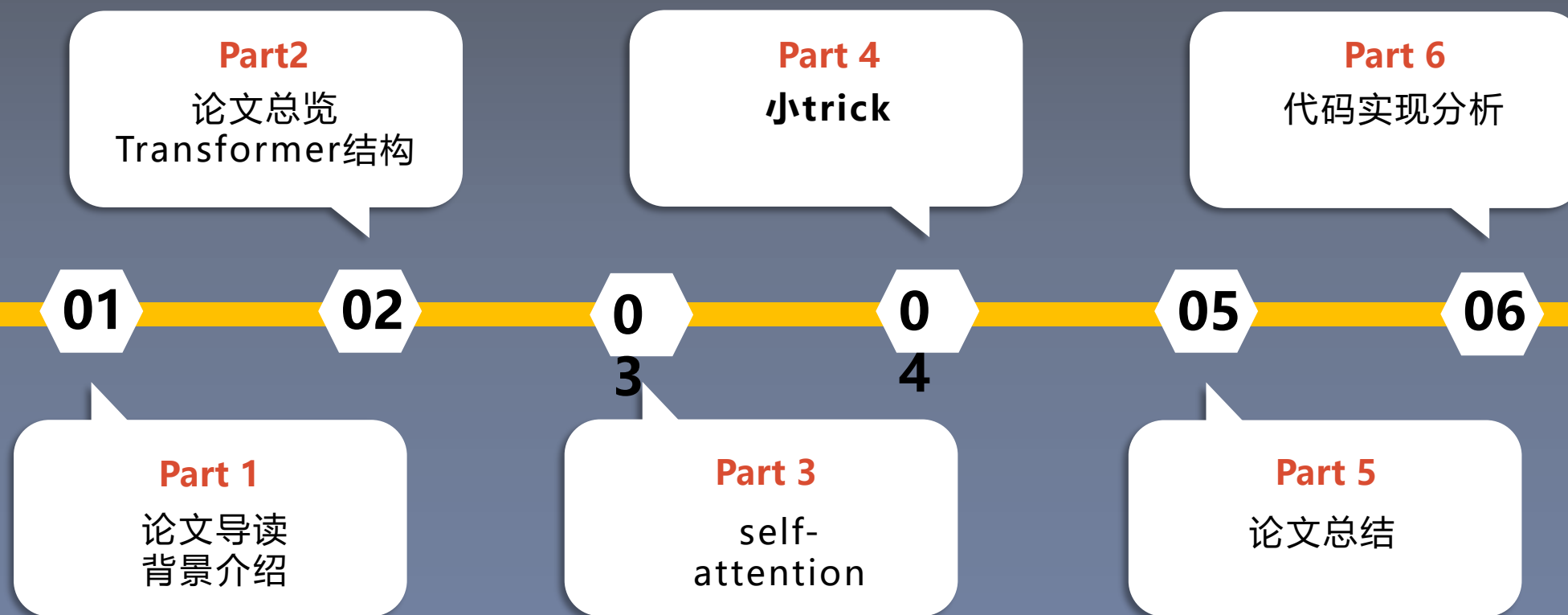
# 学习目标

Learning objectives



# 课程安排

The schedule of course





# 第一课：论文导读

The first lesson: the paper guide

---



# 目录

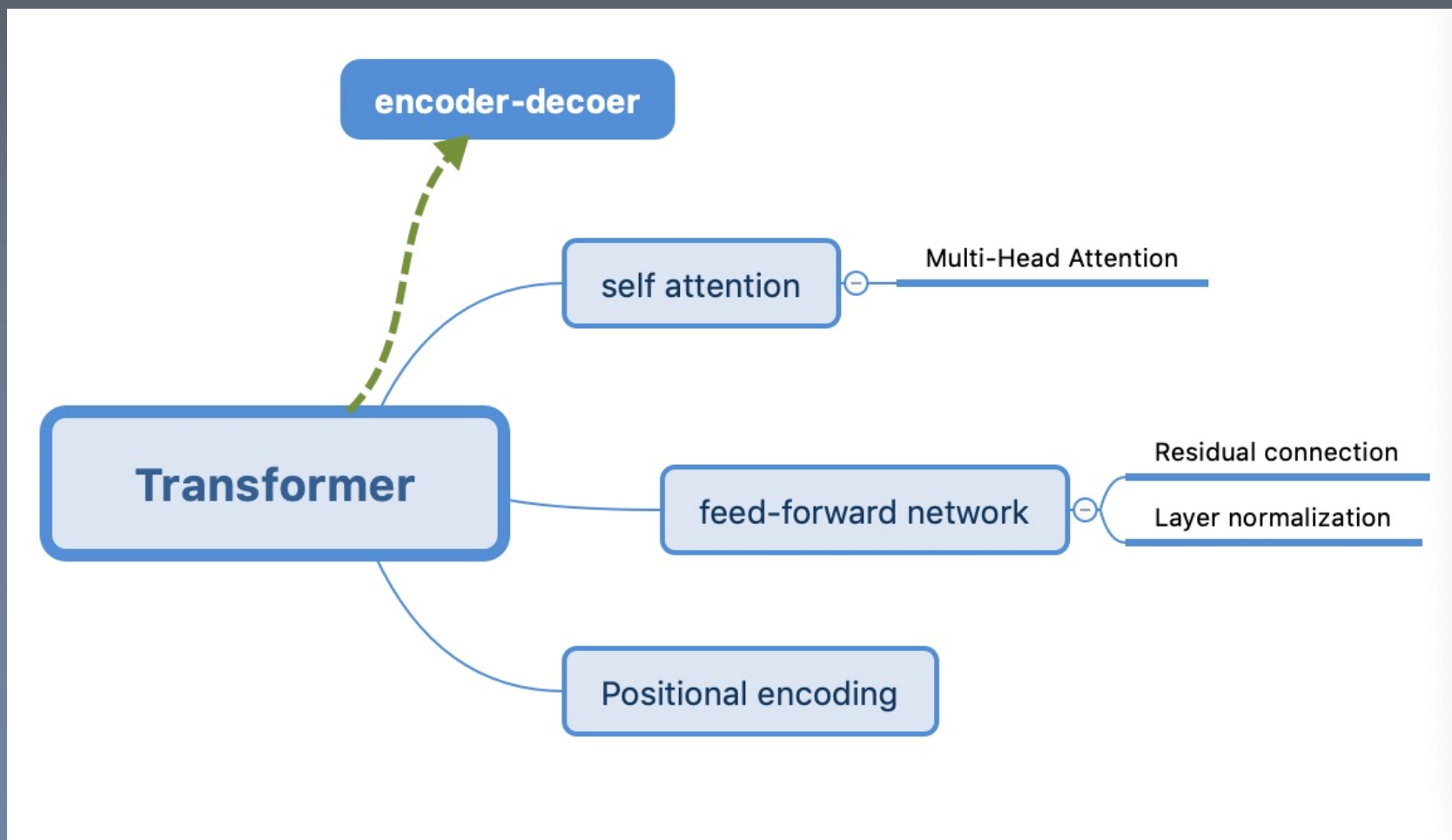
1/ 论文研究背景、成果及意义

2/ 论文泛读

3/ seq2seq以及attention回顾

3/ 本课回顾及下节预告







# 论文研究背景、成果及意义

---





# 研究背景

## Research background

eval.en x eval.de x

1 When I was 11, I remember waking up one morning to the sound of joy in my house.  
2 My father was listening to BBC News on his small, gray radio.  
3 There was a big smile on his face which was unusual then, because the news mostly depressed him.  
4 "The Taliban are gone!" my father shouted.  
5 I didn't know what it meant, but I could see that my father was very, very happy.  
6 "You can go to a real school now," he said.  
7 A morning that I will never forget.  
8 A real school.  
9 You see, I was six when the Taliban took over Afghanistan and made it illegal for girls to go to school.  
10 So for the next five years, I dressed as a boy to escort my older sister, who was no longer allowed to be at school.  
11 It was the only way we both could be educated.  
12 Each day, we took a different route so that no one would suspect where we were going.  
13 We would cover our books in grocery bags so it would seem we were just out shopping.  
14 The school was in a house, more than 100 of us packed in one small living room.  
15 It was cozy in winter but extremely hot in summer.  
16 We all knew we were risking our lives -- the teacher, the students and our parents.  
17 From time to time, the school would suddenly be canceled for a week because Taliban were suspicious.  
18 We always wondered what they knew about us.  
19 Were we being followed?  
20 Do they know where we live?  
21 We were scared, but still, school was where we wanted to be.  
22 I was very lucky to grow up in a family where education was prized and daughters were treasured.  
23 My grandfather was an extraordinary man for his time.  
24 A total maverick from a remote province of Afghanistan, he insisted that his daughter, my mom, go to school.  
25 But my educated mother became a teacher.  
26 There she is.  
27 She retired two years ago, only to turn our house into a school for girls and women in our neighborhood.

eval.en x eval.de x

1 Als ich 11 Jahre alt war, wurde ich eines Morgens von den Klängen heller Freude geweckt.  
2 Mein Vater hörte sich auf seinem kleinen, grauen Radio die Nachrichtensendung der BBC an.  
3 Er sah sehr glücklich aus, was damals ziemlich ungewöhnlich war, da ihn die Nachrichten meistens depressiv machten.  
4 Er rief: "Die Taliban sind weg!"  
5 Ich wusste nicht, was das bedeutete, aber es machte meinen Vater offensichtlich sehr, sehr glücklich.  
6 "Jetzt kannst du auf eine richtige Schule gehen," sagte er.  
7 Diesen Morgen werde ich niemals vergessen.  
8 Eine richtige Schule.  
9 Die Taliban ergriffen die Macht in Afghanistan, als ich sechs war, und verboten es Mädchen, zur Schule zu gehen.  
10 Deshalb verkleidete ich mich fünf Jahre lang als Junge und begleitete meine ältere Schwester, die nicht zur Schule gehen durfte.  
11 Nur so konnten wir beide zur Schule gehen.  
12 Jeden Tag nahmen wir einen anderen Weg, sodass niemand erraten konnte, wohin wir gingen.  
13 Wir versteckten unsere Bücher in Einkaufstüten, damit es so aussah, als würden wir nur einkaufen gehen.  
14 Unterrichtet wurden wir in einem Haus, über 100 Mädchen in einem kleinen Wohnzimmer.  
15 Im Winter war es gemütlich, aber im Sommer war es unglaublich heiß.  
16 Wir alle wussten, dass wir unser Leben riskierten: Lehrer, Schüler und unsere Eltern.  
17 Immer wieder musste der Unterricht plötzlich für eine Woche ausfallen, weil die Taliban Verdacht gegen die Schule hegte.  
18 Wir waren uns nie sicher, wie viel sie über uns wussten.  
19 Verfolgten sie uns?  
20 Wussten sie, wo wir wohnten?  
21 Wir hatten Angst, aber wir wollten trotzdem zur Schule gehen.  
22 Ich hatte großes Glück in einer Familie aufzuwachsen, in der Bildung als wichtig galt und Töchter gerne zur Schule geschickt wurden.  
23 Mein Großvater war seiner Zeit weit voraus.  
24 Ein Außenseiter aus einer entlegenen Provinz Afghanistans. Er bestand darauf, seine Tochter – meine Mutter – zur Schule zu schicken.  
25 Meine gebildete Mutter aber wurde Lehrerin.  
26 Das ist sie.  
27 Vor zwei Jahren ging sie in den Ruhestand, nur um unser Haus in eine Schule für Mädchen und Frauen umzuwandeln.  
28 Und mein Vater – hier zu sehen – war der Erste in seiner Familie, der jemals eine Schulbildung erhielt.  
29 Für ihn war stets klar, dass seine Kinder eine Ausbildung erhalten würden, auch seine Töchter, trotz der schwierigen Umstände.  
30 Er sah es als ein viel größeres Risiko an, seine Kinder nicht zur Schule zu schicken.  
31 Ich weiß noch genau, dass ich in den Jahren unter den Taliban manchmal so frustriert war von unseren Umständen.

## 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



重点 重点来了!

## 翻译效果衡量指标bleu

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

参考译文

It is a nice day today

译文

Today is a nice day

It is a nice day today

Today is a nice day

5/6

It is a nice day today

Today is a nice day

2/4



# 研究背景

Research background

## 翻译效果衡量指标bleu



深度之眼  
deepshare.net



重点 重点来了!

参考译文

the the the the

译文

The cat is standing on the ground

Count<sub>clip</sub> = min(Count, Max\_Ref\_Count)



2

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



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

# 研究成果

## Research Results



深度之眼  
deepshare.net

$n < d$

0

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]	<u>26.03</u>	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]	<u>26.36</u>	<b><u>41.29</u></b>	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$3.3 \cdot 10^{18}$	
Transformer (big)	<b><u>28.4</u></b>	<b><u>41.0</u></b>	$2.3 \cdot 10^{19}$	

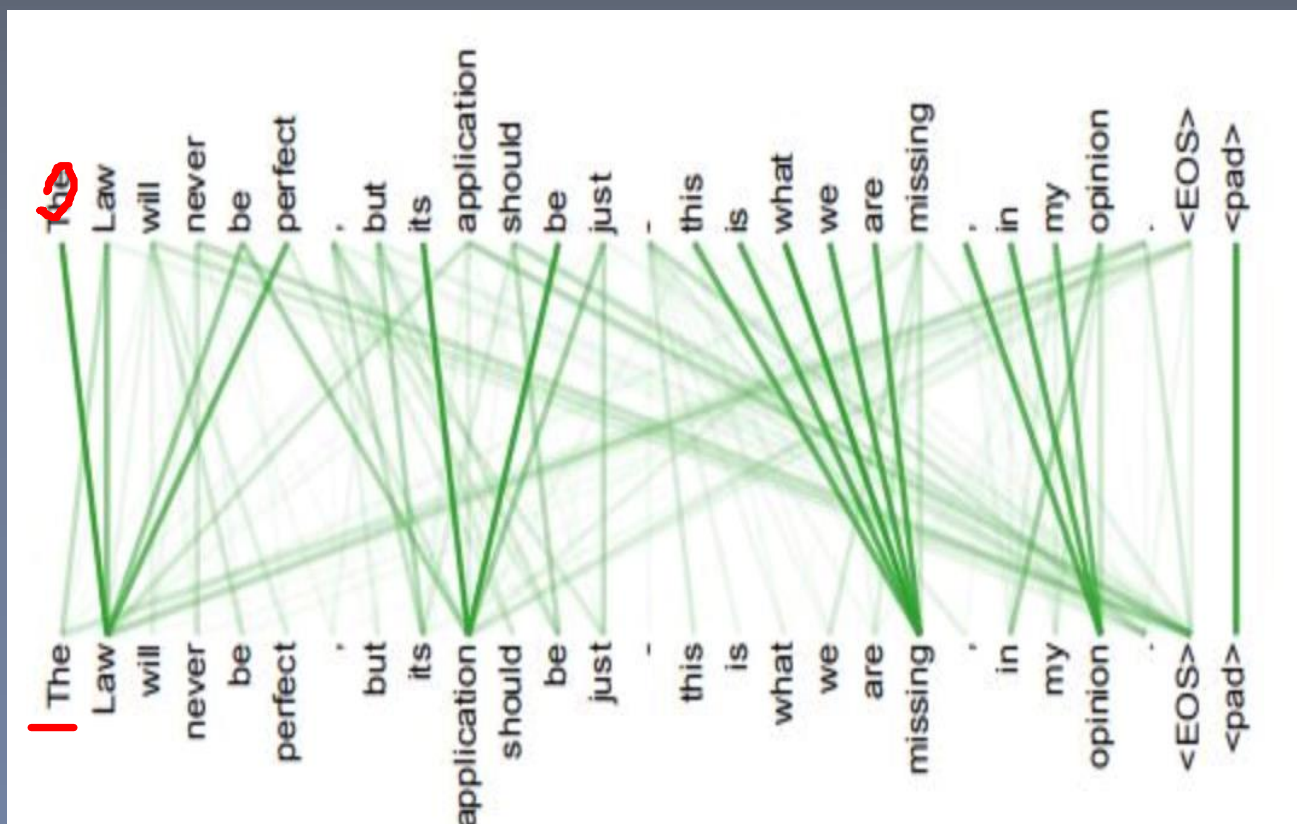
1 在WMT 2014English-to-German 翻译任务上比其它模型的bleu值高出两个点。

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。

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

# 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



# 研究意义

Research Meaning



深度之眼  
deepshare.net



重点 重点来了!

PLAN D

## 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tinybert模型。





# 论文泛读

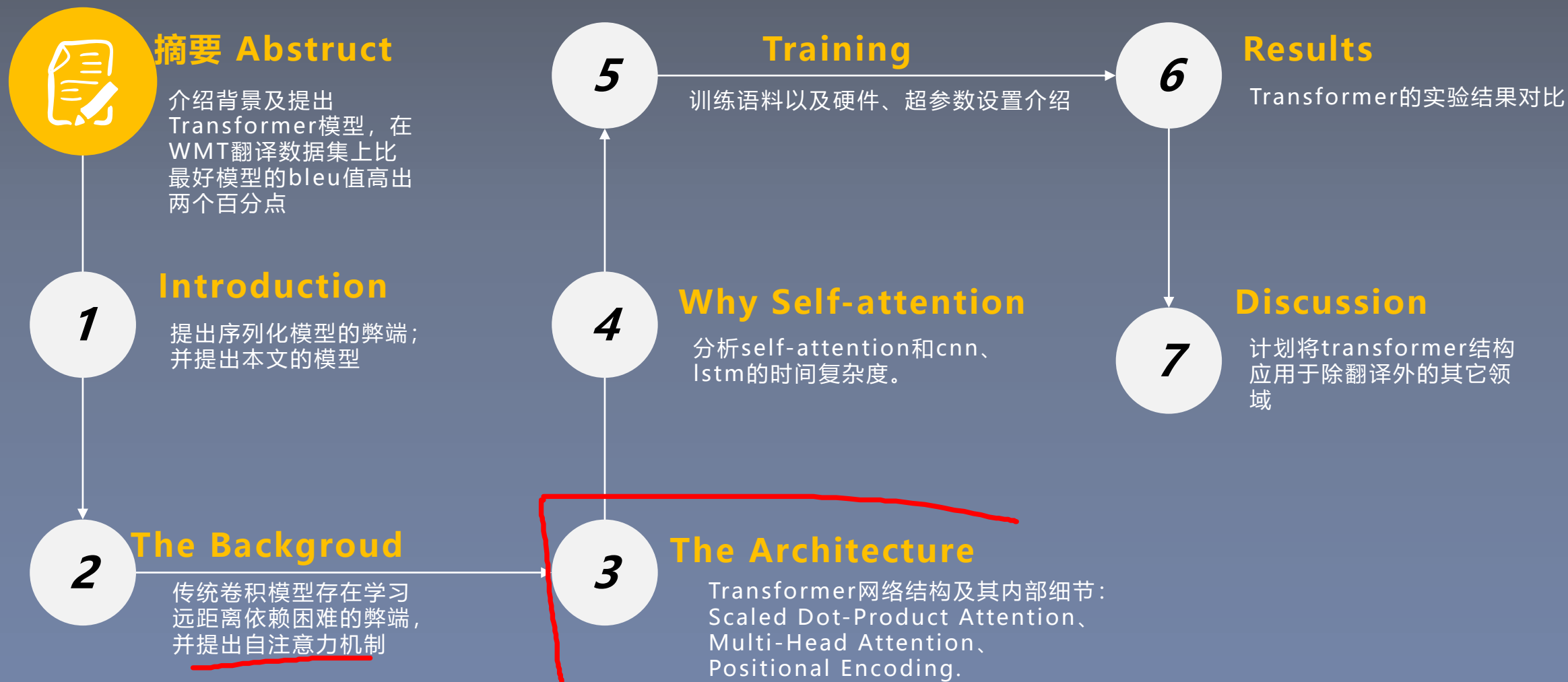
Strcuture of Paper

---



# 论文结构

## Structure of Papers



# 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

---

1. Introduction

2. Background

3. Model Architecture

3.1 Encoder and Decoder Stacks

3.2 Attention

3.2.1 Scaled Dot-Product Attention

3.2.2 Multi-Head Attention

3.2.3 Applications of Attention in our Model

3.3 Position-wise Feed-Forward Networks

3.4 Embeddings and Softmax

3.5 Positional Encoding

4. Why Self-Attention

5 Training

5.1 Training Data and Batching

5.2 Hardware and Schedule

5.3 Optimizer

5.4 Regularization

6 Results

6.1 Machine Translation

6.2 Model Variations

7 Discussion



# seq2seq以及 attention的回顾

Strcuture of Paper

---



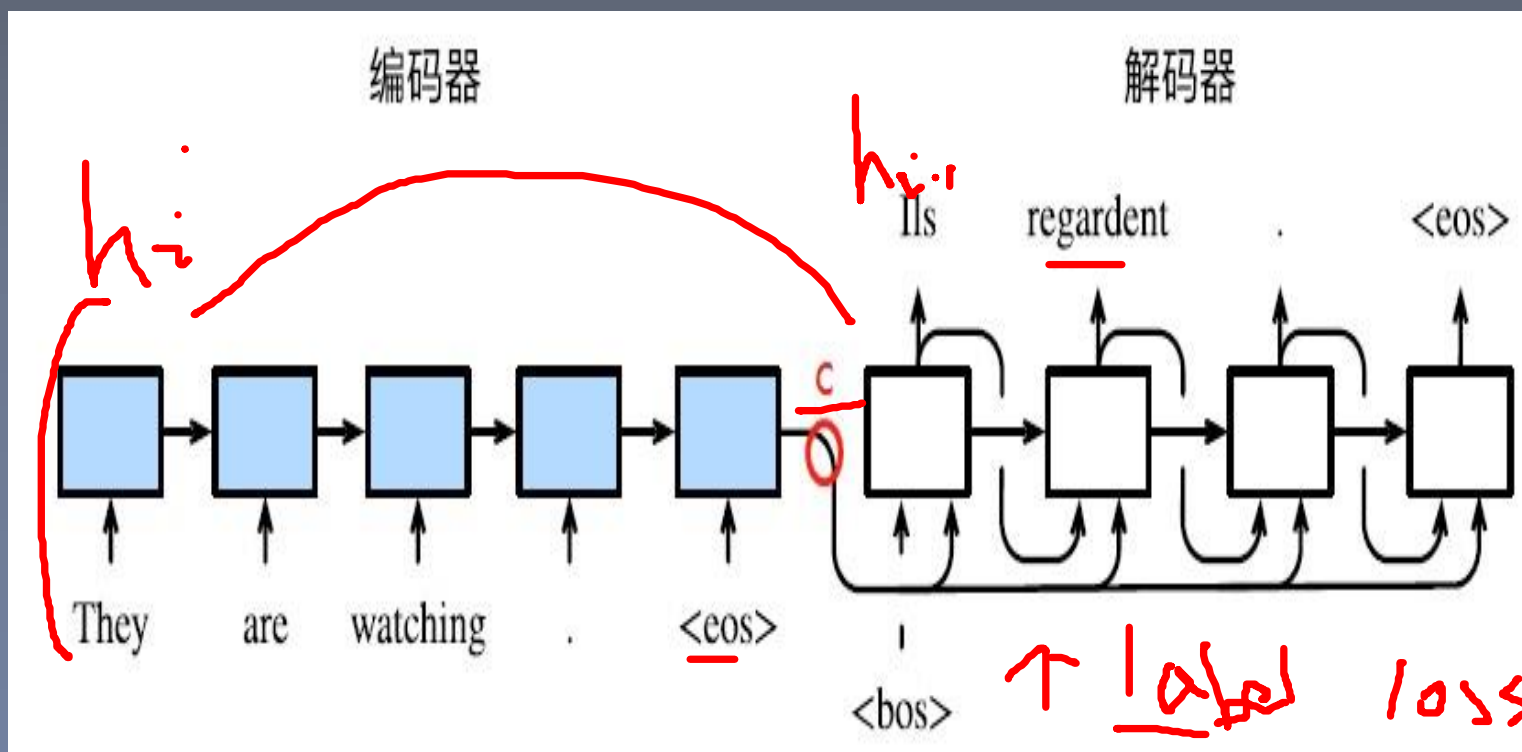
# 论文结构

Structure of Papers



深度之眼  
deepshare.net

## seq2seq模型以及attention机制



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



# 本课回顾及下节预告

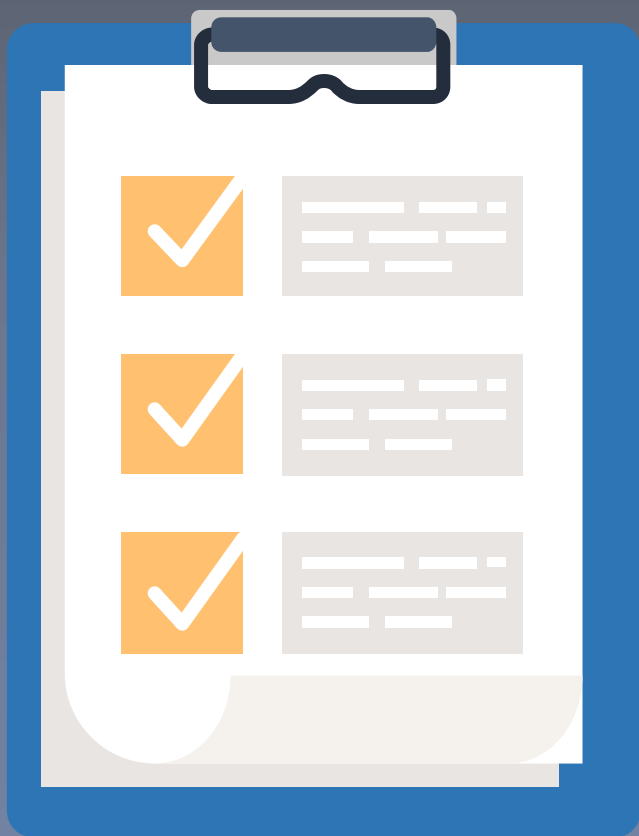
Review in the lesson and Preview of next lesson

---



# 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。

# 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

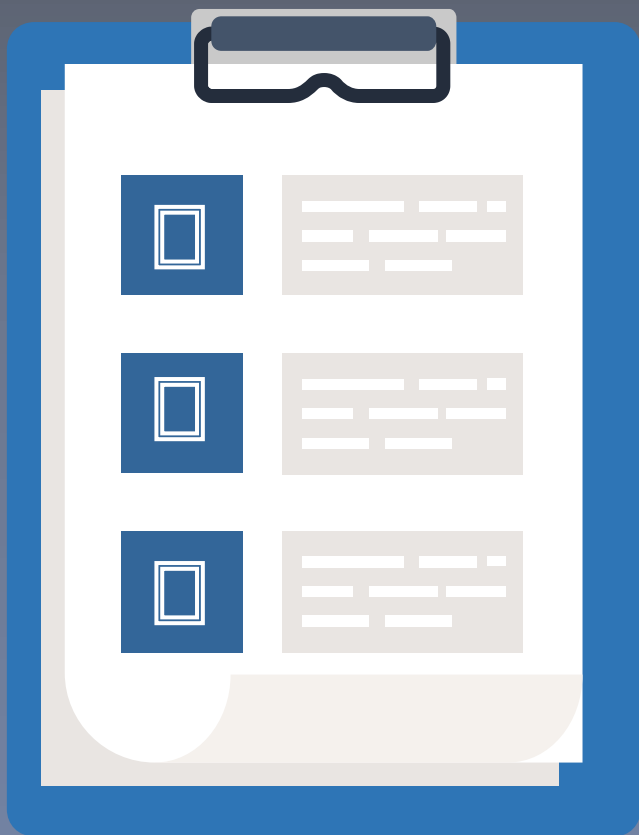
网络超参数设置，学习率，batchsize等  
实验结果分析对比

## 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



# ——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

