

Attention Is All You Need

(注意力机制就是你所需要的全部)

Ashish Vaswani · Noam Shazeer · Niki Parmar · Jakob Uszkoreit ·
Llion Jones · Aidan N. Gomez · Łukasz Kaiser · Illia Polosukhin
Google Brain / Google Research / University of Toronto
arXiv:1706.03762v7 [cs.CL]

中文译本 — Translated by LaoWang

摘要

1 引言

2 背景

3 模型架构

3.1 编码器和解码器堆叠

3.2 注意力机制

3.2.1 缩放点积注意力

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

3.2.2 多头注意力

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O$$

$$\text{head}_i = \text{Attention}(Q W_i^Q, K W_i^K, V W_i^V)$$

3.2.3 注意力在我们模型中的应用

3.3 逐位置前馈网络

$$\text{FFN}(x) = \max(0, xW_1 + b_1) W_2 + b_2$$

3.4 嵌入与 Softmax

3.5 位置编码

$$\text{PE}(\text{pos}, 2i) = \sin(\text{pos} / 10000^{(2i/d_{\text{model}})})$$

$$\text{PE}(\text{pos}, 2i+1) = \cos(\text{pos} / 10000^{(2i/d_{\text{model}})})$$

4 为什么用自注意力

5 训练

5.1 训练数据与批处理

5.2 硬件与训练时间

5.3 优化器

```
lrate = d_model^(-0.5) * min(step_num^(-0.5), step_num * warmup_steps^(-1.5))
```

5.4 正则化

6 结果

6.1 机器翻译

6.2 模型变体

6.3 英语成分句法解析

7 结论

— 全文完 —

译注：本译文保留所有模型名、架构名等专有名词的英文原文，仅翻译解释性文字。公式和引用编号与原文一致。Figure 和 Table 请参考原论文。